THE QUANTITATIVE STUDY OF CHANGES IN ANATOMY

by

Jacob D. Hinkle

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Bioengineering

The University of Utah

December 2015

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Jacob D. Hinkle**

has been approved by the following supervisory committee members:

Sarang C. Joshi	, Chair	**7/30/2013** Date Approved
Edward V. R. DiBella	, Member	**7/30/2013** Date Approved
P. Thomas Fletcher	, Member	**7/30/2013** Date Approved
Robert S. Macleod	, Member	**7/30/2013** Date Approved
Billie Jean Salter Jr.	, Member	**7/30/2013** Date Approved

and by **Patrick A. Tresco** , Chair/Dean of

the Department/College/School of **Bioengineering**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

The statistical study of anatomy is one of the primary focuses of medical image analysis. It is well-established that the appropriate mathematical settings for such analyses are Riemannian manifolds and Lie group actions. Statistically defined atlases, in which a mean anatomical image is computed from a collection of static three-dimensional (3D) scans, have become commonplace. Within the past few decades, these efforts, which constitute the field of computational anatomy, have seen great success in enabling quantitative analysis. However, most of the analysis within computational anatomy has focused on collections of static images in population studies. The recent emergence of large-scale longitudinal imaging studies and four-dimensional (4D) imaging technology presents new opportunities for studying dynamic anatomical processes such as motion, growth, and degeneration. In order to make use of this new data, it is imperative that computational anatomy be extended with methods for the statistical analysis of longitudinal and dynamic medical imaging.

In this dissertation, the deformable template framework is used for the development of 4D statistical shape analysis, with applications in motion analysis for individualized medicine and the study of growth and disease progression.

A new method for estimating organ motion directly from raw imaging data is introduced and tested extensively. Polynomial regression, the staple of curve regression in Euclidean spaces, is extended to the setting of Riemannian manifolds. This polynomial regression framework enables rigorous statistical analysis of longitudinal imaging data. Finally, a new diffeomorphic model of irrotational shape change is presented. This new model presents striking practical advantages over standard diffeomorphic methods, while the study of this new space promises to illuminate aspects of the structure of the diffeomorphism group.

Mathematics is a part of physics. Physics is an experimental science, a part of natural science. Mathematics is the part of physics where experiments are cheap.

– Vladimir I. Arnold

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

With the invention of modern three-dimensional (3D) medical imaging technologies like computed tomography (CT) and magnetic resonance imaging (MRI), clinicians gained the ability to inspect *in vivo* internal patient anatomy much more accurately and naturally than with traditional x-ray imaging. These three-dimensional modalities have proven useful for diagnosis and treatment planning, relying chiefly on 3D visualization and expert interpretation by trained clinicians. Qualitative analyses are very informative, but 3D imaging has promised from the beginning to also enable more quantitative analysis. Within the last twenty years, substantial effort has been made to develop these quantitative analyses using anatomical information contained within 3D images. For example, radiation therapy requires accurate quantitative demarcation of tumors using CT images to plan and execute precise treatment [12, 16]. In this dissertation, I develop new modeling techniques for studying shape data acquired using 3D and four-dimensional (4D) medical imaging.

The subject of this work is the analysis of *anatomical shapes*: commonly tissues, organs, or other geometric regions within the human body. In addition to simple measures of shape such as size and position, complex descriptors of shape are of significant interest when comparing multiple anatomical shapes. For instance, the corpus callosum of an Alzheimer's disease patient is expected to be thinner than that of a patient without the disease [11]. The relative thickness of a shape, particularly in specific regions within a structure such as the corpus callosum, is a complex property that is not easily described with simple measurements, as are size and position. These detailed descriptions of shape, when studied in the presence of various medical conditions, give insight into the nature of diseases and provide biomarkers which aid diagnosis and treatment planning [19, 46, 37].

The primary focus of this dissertation is temporal *shape change*, as opposed to analysis of static 3D shapes. The study of shape change is relevant in a broad range of clinical applications, from modeling respiratory motion of tumors (on time scales of a few seconds) to the growth of brain structures during neurodevelopment and neurodegeneration due to

aging (each on the time scale of years or decades). A common thread is that medical image data are acquired at multiple times, from which transformations representing continuous temporal changes in anatomy are computed.

1.1 Deformable Templates

Throughout the development of modern shape analysis, the idea that spatial transformations are the principal objects of interest in anatomy has been pervasive. As long ago as 1917, D'Arcy Thompson [45] recognized the usefulness of studying anatomy of species using a representative template anatomy, upon which are defined anatomical coordinates. He described individuals by transforming these standard coordinates to match the individual's anatomy.

That transformations are informative objects is a fundamental notion [46, 37, 1]. Take as an example a swinging pendulum which is photographed repeatedly. Each individual photograph shows a pendulum in some configuration, from which one might infer the size, shape, and physical construction of the pendulum. However, by modeling the swinging depicted in the series of images, information is acquired about the modes of motion and, ultimately, the physical laws determining that motion. In analogy to Thompson's work, the goal is to gain knowledge about *fundamental* processes such as gravitation and Newton's laws of motion using image data, instead of merely improving the imaging of a single pendulum.

D'Arcy Thompson focused on conformal (angle-preserving) transformations to represent growth and evolution of anatomical shapes [45]. The foundational work of Amit, Grenander, and Piccioni [1] initiated the theory of *deformable templates*, which realizes the ideal of studying quantitatively how a template anatomy deforms to generate the space of anatomies for a population or species. Clearly they shared Thompson's perspective that statistical analysis of *deformations* is the critical objective, as opposed to the direct statical analysis of individual medical images. This perspective persists today within the medical image analysis community, with modern progress coming in the form of new deformation models and new statistical methods accompanying these more complex models [28, 26, 40, 8, 10, 33, 47].

1.2 Shape Transformations

1.2.1 Low-Dimensional Transformations

Among the simplest transformations are translations, obtained by performing a constant spatial shift to the image, and global rotations. Combinations of translations and rotations

are referred to as *rigid* transformations. While useful for modeling coarse motion of simple shapes, rigid transformations are often considered a nuisance, as they usually represent the relative position of the subject to the scanner rather than interesting biological phenomena [4, 29].

Shapes are also commonly transformed by *scaling*, in which they undergo a global contraction or expansion. In contrast to rigid transformations which represent only the pose of an object, scale represents an anatomical property that is of interest in certain studies [32]. The scale of anatomical structures is often examined in growth and aging studies. Scale is commonly studied due to its simplicity and because it is related to easily-measurable properties such as shape volume.

Scaling, translation, rotation, as well as other global transformations that allow anisotropic stretching and shearing (so-called affine transformations), are referred to as low-dimensional transformations since they are described using a small number degrees of freedom. These transformations are simple to compute and, due to the low degrees of freedom, are amenable to efficient computation [34, 39].

Low-dimensional transformations are useful for studying very simple shapes, such as triangles. For example, any two congruent triangles in 2D differ by a rigid transformation. Similar triangles, on the other hand, differ by a combination of a rigid transformation and a scaling. In fact, such combinations of rigid transformations and scalings are called *similarity transforms* precisely because of this fact. Key to this sort of analysis is that the *simplicity* of the shapes (triangles are described by only three points) is intimately tied to the simplicity of the low-dimensional transformations used to study them.

1.2.2 Deformable Shape Models

Low-dimensional shape transformations are useful due mainly to their simplicity and ease of implementation. However, because of their global nature, they are often not flexible enough to sufficiently describe complex shapes and images containing fine detail. As a result, within medical shape analysis, they are often only useful for coarse alignment of shape data. In order to better describe complex shapes within the deformable template framework, more flexible deformation models are necessary [42, 7, 9].

Many approaches have been taken to develop more flexible deformation models [1, 6, 36, 31, 41]. These more flexible models generally define a transformation locally at every point x in space, instead of globally. In general, transformations are determined by where each point x in the image domain is mapped; the result is a *displacement* vector field $u(x)$ representing the deformation $x \mapsto x + u(x)$. Considering all such mappings from the image

domain to itself leads to highly ill-posed deformation estimation problems, often having nonsensical solutions. The problem is that although such a wide class of mappings is able to provide excellent shape matching, the solutions are highly irregular, with neighboring pixels potentially being mapped to different regions. In order to regularize these problems, deformations are placed within some restricted (but still highly flexible) family.

A simple regularization scheme is to estimate $u(x)$ directly, while penalizing nonsmooth displacement fields. This is the basis of the elastic deformation model [6]. Another common approach is to represent displacements within some low-dimensional vector space using smooth *basis functions*, as is common in spline modeling [42, 41]. These methods regularize deformation estimation problems by guaranteeing smooth solutions, but cannot guarantee that the resulting deformations will be invertible.

Invertibility is an important requirement when determining anatomical correspondences in which it is expected that each point in a given image matches one and only one point in the other image. In the case of organ motion, for instance, a single point can move to one and only one point at a future time. The desire for more realistic smooth deformations led to the theory of *diffeomorphic* image transformations (those which are both smooth and invertible) [7, 27]. Diffeomorphic transformations are obtained by composing many small smooth deformations with one another. The small displacements approximate smooth *velocity fields* along which the image domain *flows*, just as a physical fluid's motion is described by a flow along velocity vector fields. This fluid model of diffeomorphisms is the essential element of the large deformation diffeomorphic metric mapping (LDDMM) framework, and constitutes a solid connection between medical image analysis and fluid dynamics [26, 35].

1.3 Lie Group Actions and Computational Anatomy

The examples of low-dimensional transformation models, as well as diffeomorphic fluid deformation models, share a common setting: that of Lie group actions. Informally, Lie groups are sets of transformations which may be composed with one another, the collection of which forms a smooth space (a manifold). These groups act on shapes and images by transforming the underlying spatial domain. The same group, such as the group of rotations $\mathrm{SO}(n)$ or the full diffeomorphism group $\mathrm{Diff}(\mathbb{R}^n)$, generally acts on many types of objects:[1] images, point-sets describing canonically-defined anatomical landmarks, surfaces generated

[1] I mention the *full* diffeomorphism group, indicating that it contains other previously mentioned groups. Indeed, all of the low-dimensional transformations are found within subgroups of $\mathrm{Diff}(\mathbb{R}^n)$, a fact which implies that diffeomorphisms are the most general way of studying smooth invertible shape transformations.

by segmentation, or other more exotic objects such as tensor fields or distributions. Most of the applications in this work use Lie groups acting on continuously defined images, from which other shape structures are commonly derived.

Lie group actions represent "continuous symmetry" and provide a unifying framework in which to view modern medical shape analysis [37, 19, 35]. Lie groups are particularly interesting as they allow us to use their manifold structure to define *metrics* between transformations.[2] This provides a setting for formal statistical analysis of *transformations*, as opposed to images or other shape descriptors [28]. This method of performing statistics using distance metrics defined in terms of Lie group actions characterizes the framework commmonly referred to as *computational anatomy*, an umbrella term that encompasses many of the methods presented in this dissertation.

The group structure of Lie groups is also useful in separating useful and nuisance transformations. For instance, when studying landmark point sets, it is common to ignore the pose and size (determined by similarity transform) of the data and focus on more interesting modes of shape variability. Using Lie groups, equivalence of landmark point sets under similarity transforms is formalized, leading directly to a natural shape space in which pose and size are unable to influence analysis. This space, Kendall's shape space [30, 32], elegantly captures the informative parts of shape variability while avoiding the influence of irrelevant pose and size information. This is just one example of how Lie group actions facilitate different treatment of transformation subgroups representing different physical phenomena.

1.3.1 Statistical Shape Analysis

The desire for more accurate modeling of complex shapes spurred the development of the more flexible deformation models already mentioned. Interestingly, certain transformation models (LDDMM in particular) provide not only a method for deforming shapes and images, but a definition of the *size* of a transformation [28, 3]. In fact, in the case of LDDMM, the size of a deformation also defines a metric on the group of diffeomorphisms of the image domain. Such spaces equipped with metrics are particularly convenient settings in which to do statistics.

---

[2]Note that the term *metric* in this context refers to an always positive, symmetric distance function $d : M \times M \to \mathbb{R}^+$ between pairs of transformations, satisfying the triangle inequality.

1.3.1.1 Atlas Building

Commonly, imaging studies are performed in which images are collected for a number of subjects, and statistics are computed in order to describe the population. For instance, summary measures such as the volumes of anatomical structures are commonly computed and averaged across a population. A particularly useful application of statistical shape analysis is the computation of an "average" of a sample of images. The metrics of the LDDMM framework are instrumental in this context.

This method of shape averaging, usually called *atlas building*, using LDDMM is as follows. An average, or *template*, image is defined as an image which minimizes the "sum of squared distances" to the data [28].[3] Under this definition, the template is a Fréchet mean of the sample images, using a metric on images induced by the metric on diffeomorphisms. Thus, atlas building generalizes the notion of sample mean from Euclidean spaces to images, using the physically meaningful geometry of the transformation space, as opposed to the natural vector space structure of images.

As a simple example, consider computing the average of the two images shown in the top row of Fig. 1.1. Using the vector space structure of images, the mean image would be computed by simply averaging image intensities at every pixel location. This results in an image that is a blended mix of the input images, and that does not represent a crisp circular structure. By contrast, building an *atlas* by computing the Fréchet mean image using a suitable metric on the diffeomorphism group gives rise to a template image representing a circle whose radius is between those in the input images. In the context of medical shape analysis, this result is more realistic, as it corresponds to a notion of averaging the embedded shapes within images while preserving their topology and smooth structure. This example serves to illustrate the fundamental principle of computational anatomy: that statistical analysis within Lie group actions representing physically realistic transformations leads to physically interpretable results.

1.3.2 Regression Analysis and Curve-Fitting

Atlas-building provides a natural method for computing mean anatomical shapes. Principal geodesic analysis [15] provides an extension of principal component analysis to the manifold setting, which is analogous to computation of sample covariance for manifold-valued shape data. These measures are extremely useful, but, typically, more information

[3]Note that here the template image is defined, but given nonimage data, the definition of a general template object such as a template surface is a straightforward adaptation.

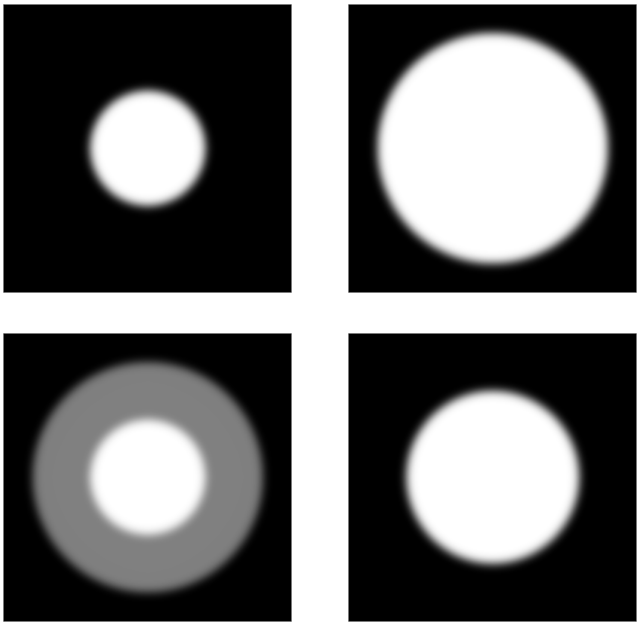


Figure 1.1. Example of shape averaging. The two input images (top row) contain circles of different radii. The pixel-wise average (lower left) is not an image containing a sharp circle. However, using deformation distance, the Fréchet mean of these images contains a sharp circle (lower right).

is available for each parcel of imaging data. For instance, images are always accompanied by study metadata such as time, age, weight, relevant clinical test scores, etc. In such cases, links between shapes and these scalar clinical variables are of special interest.

Given multiple types of "metadata," one approach to finding relationships is to combine these variables with an atlas and residual deformations, then perform a post-hoc analysis [43]. Such methods are able to detect common patterns in the data, relating combinations of the clinical variables and aspects of the shape data, without making ad-hoc assumptions that the relationships take a particular form. The resulting relationships between observed shape and clinical variables constitute imaging-based biomarkers, which augment or replace clinical test scores that are often obtained through expensive or painful methods.

Temporal metadata such as patient age and time are of particular interest. Changes in shape with respect to these variables represent long-scale and short-scale *shape change*, corresponding to growth or aging at long scales and organ motion at short scales. Instead of a post-hoc analysis, in order to study shape change, one makes the ad-hoc assumption that the shape of an individual organ is determined by the time variable. Given a sample of imaging data, collected at various times or ages, a temporally varying image representing a deforming *base image* is then estimated. Optimal curve estimation in this sense is generally

referred to as either *curve regression* (when viewed in a statistical context) or *curve fitting*.

Methods for fitting curves of diffeomorphisms to image and shape data have evolved in the past decade [10, 13, 38]. Consistent with the philosophy of computational anatomy, modern approaches to these *curve regression* problems leverage the metric structure of the diffeomorphism group. For instance, in nonparametric kernel regression, points on the regression curve are defined as a weighted Fréchet mean, using a kernel to provide smoothly time-varying weights [10]. Other methods phrase nonparametric curve regression as a variational problem, using a regularization which penalizes rapid changes in curve direction [17]. Recently developed methods approach shape regression using parametric families of curves instead of regularization schemes on the unparametrized family of all curves. The most widely-known of these are geodesic regression methods which fit a single geodesic through a collection of observed data [13, 38]. The resulting curves are very useful in statistical studies due to their compact representation, but lack the flexibility to accurately match shape data in many cases.

1.4 Main Contributions

The achievements presented in this dissertation lay the groundwork for the expansion of computational anatomy from the analysis of static 3D images to 4D modeling for individualized medicine and longitudinal statistical studies of shape. Specific contributions on this front include an extension of image-based diffeomorphic motion modelling to a new approach which estimates diffeomorphic motion from raw imaging data, while simultaneously reconstructing a deforming template image. In addition to this advance in 4D shape modelling, longitudinal analysis is propelled in this dissertation by the use of a new parametric family of regression curves, extending the benefits of recently developed geodesic models to provide more flexibility.

In addition to new models of shape change within established diffeomorphic shape models, this dissertation includes a contribution toward better understanding the structure of the diffeomorphism group itself. As discussed above, deformation models have been expanded and refined throughout the development of deformable template theory. From low-dimensional transformations to smooth linear displacement models, culminating in modern diffeomorphic deformation models, exploration of these spaces has led to new insights and capabilities. Although the theory of incompressible diffeomorphisms has been well-studied in the context of fluid dynamics [2], other subspaces of the diffeomorphism group have seen little attention. In this dissertation, I explore a particular space which is

the irrotational counterpart to the incompressible subgroup, work that is closely related to very recent efforts in the applied mathematics and fluid dynamics communities.

The work presented in this dissertation can be summarized by the following major contributions:

Chapter 2: A four-dimensional (4D) image reconstruction method is presented, in which a diffeomorphic motion model is used to estimate organ motion using raw projection data while simultaneously estimating a deforming base image. Results are presented for conebeam and fanbeam CT, in phantom studies as well as on patient data, validating the accuracy of the obtained motion estimates. As part of this framework, I derive a method of enforcing an incompressibility constraint, globally or locally, during motion estimation.

This work was published in the included manuscript Hinkle et al. [21] and is supported by conference proceedings Hinkle et al. [22], Hinkle et al. [24], and Hinkle et al. [23], as well as the related papers Geneser et al. [18] and Szegedi et al. [44].

Chapter 3: A related application, in which motion estimation is used to correct the pose information during 3D image reconstruction from an uncalibrated imaging device, is presented. Whereas in 4D image reconstruction, the scanner geometry is known and anatomical motion is estimated, in this application, the situation is reversed. The anatomy is presumed to be static, while the scanner geometry is dynamic and not well-calibrated.

Chapter 4: The established geodesic regression method [14, 38] is extended by introducing a framework for fitting higher-order polynomials on Riemannian manifolds and Lie groups. These more flexible classes of curves enable more accurate curve fitting, while maintaining a compact representation.

This chapter includes a reprint of a manuscript, currently accepted to the Journal of Mathematical Imaging and Vision (Springer), which is an extension of the conference proceeding Hinkle et al. [25].

Chapter 5: A new space of *irrotational* diffeomorphisms, called IDiff, is introduced. The geometry of this space enables extremely efficient image registration and atlas-building algorithms. The deformations in IDiff are determined chiefly by local expansion and contraction. In light of the conventional study of expansion and contraction in neurodevelopment studies, irrotational diffeomorphisms are potentially a more realistic

model for growth and aging of brain structures. In addition, the exploration of this space in real applications gives new insights into the structure of the diffeomorphism group.

This chapter constitutes an extension of the conference proceeding Hinkle and Joshi [20], and will be the basis for a future journal publication.

Appendix A: This appendix includes a thorough treatment of Lie groups with left and right invariant metrics. The adjoint representation is presented and connected to the inner automorphism group. Invariance of vector fields and metrics is defined, and covariant differentiation is discussed in these settings. The Euler-Poincaré equation (the basis for the famous EPDiff equation) is derived for both right and left invariant metrics. Most interestingly, little-known formulas derived in Bullo [5] are duplicated, providing a very elegant treatment of Jacobi fields in Lie groups with left and right invariant metrics.

Appendix B: This appendix includes a detailed treatment of the modern theory of computational anatomy, wherein objects are transformed using Lie group actions. The theory of infinitesimal generators and momentum maps connects dynamics in object space to dynamics in the group. Furthermore, momentum maps provide a natural example of a *conserved property* under geodesic evolution. This abstract theory presents the most modern mathematical framework encompassing many kinds of shape analysis. Finally, the most fundamental application, the diffeomorphism group acting on scalar images, is presented as an example.

Appendix C: This appendix comprises a reprint of the publication Geneser et al. [18]. That publication is a novel application of the 4D MAP image reconstruction algorithm developed in Chapter 2, in which motion modelling via 4D image reconstruction is used in conjunction with stochastic models of breathing to estimate a *distribution* of doses delivered using a static radiation therapy plan.

Chapter 6 provides a discussion of these chapters, with a summary of the work and an outlook for future work. Each of these chapters is self-contained. As a result, each chapter includes a bibliography for the references contained therein.

References

[1] Yali Amit, Ulf Grenander, and Mauro Piccioni. “Structural Image Restoration Through Deformable Templates”. In: *J. Am. Statistical Assn.* 86.414 (June 1991), pp. 376–387.

[2] Vladimir I Arnol'd. "Sur la Géométrie Différentielle des Groupes de Lie de Dimension Infinie et ses Applications à l'Hydrodynamique des Fluides Parfaits". In: *Ann. Inst. Fourier* 16 (1966), pp. 319–361.

[3] M Faisal Beg et al. "Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms". In: *International Journal of Computer Vision* 61.2 (2005), pp. 139–157.

[4] Fred L Bookstein. "Size and Shape Spaces for Landmark Data in Two Dimensions". In: *Statistical Science* (1986), pp. 181–222.

[5] Francesco Bullo. "Invariant Affine Connections and Controllability On Lie Groups". In: (1995).

[6] Gary E Christensen, Richard D Rabbitt, and Michael I Miller. "3D Brain Mapping Using a Deformable Neuroanatomy". In: *Physics in Medicine and Biology* 39.3 (1994), p. 609.

[7] Gary E. Christensen, Richard D. Rabbitt, and Michael I. Miller. "Deformable Templates Using Large Deformation Kinematics". In: *IEEE Trans. Imag. Proc.* 5.10 (Oct. 1996), pp. 1435–1447.

[8] Colin J Cotter, Darryl D Holm, et al. "Singular Solutions, Momentum Maps and Computational Anatomy". In: *1st MICCAI Workshop on Mathematical Foundations of Computational Anatomy: Geometrical, Statistical and Registration Methods for Modeling Biological Shape Variability*. 2006, pp. 18–28.

[9] Christos Davatzikos. "Nonlinear Registration of Brain Images Using Deformable Models". In: *Mathematical Methods in Biomedical Image Analysis, 1996., Proceedings of the Workshop on*. IEEE. 1996, pp. 94–103.

[10] Bradley C. Davis et al. "Population Shape Regression From Random Design Data". In: *Proceedings of International Conference on Computer Vision*. 2007.

[11] Naomi R Driesen and Naftali Raz. "The Influence of Sex, Age, and Handedness on Corpus Callosum Morphology: A Meta-Analysis." In: *Psychobiology* (1995).

[12] Jake van Dyk. *The Modern Technology of Radiation Oncology*. Medical Physics Publ., 1999.

[13] P Thomas Fletcher. "Geodesic Regression and the Theory of Least Squares on Riemannian Manifolds". In: *International Journal of Computer Vision* (2012), pp. 1–15.

[14] P. Thomas Fletcher. "Geodesic Regression on Riemannian Manifolds". In: *International Workshop on Mathematical Foundations of Computational Anatomy MFCA*. 2011.

[15] P. Thomas Fletcher et al. "Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape". In: *IEEE Trans. Med. Imag.* 23.8 (2004), pp. 995–1005.

[16] Mark Foskey et al. "Large Deformation Three-Dimensional Image Registration in Image-Guided Radiation Therapy". In: *Phys. Med. Biol.* 50 (Dec. 2005), pp. 5869–5892.

[17] François Gay-Balmaz et al. "Invariant Higher-Order Variational Problems II". In: *J. Nonlin. Sci.* 22.4 (2012), pp. 553–597.

[18] Sarah E Geneser et al. "Quantifying Variability in Radiation Dose Due to Respiratory-Induced Tumor Motion". In: *Medical Image Analysis* 15 (2011), pp. 640–649.

[19] Ulf Grenander and Michael I Miller. "Computational Anatomy: An Emerging Discipline". In: *Quarterly of Applied Mathematics* 56.4 (1998), pp. 617–694.

[20] Jacob Hinkle and Sarang Joshi. "IDiff: Irrotational Diffeomorphisms for Computational Anatomy". In: *Information Processing in Medical Imaging (IPMI)*. Springer. 2013, pp. 754–765.

[21] Jacob Hinkle et al. "4D CT Image Reconstruction with Diffeomorphic Motion Model". In: *Medical Image Analysis* 16.6 (2012), pp. 1307–1316.

[22] Jacob Hinkle et al. "4D MAP Image Reconstruction Incorporating Organ Motion". In: *Information Processing in Medical Imaging (IPMI)*. Springer. 2009, pp. 676–687.

[23] Jacob Hinkle et al. "4D MAP MRI Image Reconstruction". In: *International Conference on Computer Vision Theory and Applications (VISAPP)*. Angers, France, May 2010.

[24] Jacob Hinkle et al. "Development and Testing of a Novel, 4D Maximum a Posteriori (MAP) Image Reconstruction Algorithm". In: *American Association of Physicists in Medicine (AAPM)*. Anaheim, California, July 2009.

[25] Jacob Hinkle et al. "Polynomial Regression on Riemannian Manifolds". In: *European Conference on Computer Vision ECCV*. Florence, Italy, 2012, pp. 1–14.

[26] Darryl D Holm et al. "Soliton Dynamics in Computational Anatomy". In: *NeuroImage* 23 (2004), S170–S178.

[27] Sarang C Joshi and Michael I Miller. "Landmark Matching via Large Deformation Diffeomorphisms". In: *Image Processing, IEEE Transactions on* 9.8 (2000), pp. 1357–1370.

[28] Sarang Joshi et al. "Unbiased Diffeomorphic Atlas Construction for Computational Anatomy". In: *NeuroImage* 23 (2004), S151–S160.

[29] David G. Kendall. "A Survey of the Statistical Theory of Shape". In: *Statistical Science* 4.2 (1989), pp. 87–99.

[30] David G. Kendall. "Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces". In: *Bull. London Math. Soc.* 16.2 (1984), pp. 81–121.

[31] Peter J Kostelec, John B Weaver, and Dennis M Healy Jr. "Multiresolution Elastic Image Registration". In: *Medical Physics* 25 (1998), p. 1593.

[32] Huiling Le and David G. Kendall. "The Riemannian Structure of Euclidean Shape Spaces: A Novel Environment for Statistics". In: *Ann. Statist.* 21.3 (1993), pp. 1225–1271.

[33] Jun Ma et al. "Bayesian Template Estimation in Computational Anatomy". In: *NeuroImage* 42.1 (2008), pp. 252–261.

[34] JB Maintz and Max A Viergever. "A Survey of Medical Image Registration". In: *Medical Image Analysis* 2.1 (1998), pp. 1–36.

[35] Michael I Miller, Alain Trouvé, and Laurent Younes. "On the Metrics and Euler-Lagrange Equations of Computational Anatomy". In: *Annual Review of Biomedical Engineering* 4.1 (2002), pp. 375–405.

[36] Michael I Miller et al. "Mathematical Textbook of Deformable Neuroanatomies". In: *Proceedings of the National Academy of Sciences* 90.24 (1993), pp. 11944–11948.

[37] Michael Miller et al. "Statistical Methods in Computational Anatomy". In: *Statistical Methods in Medical Research* 6.3 (1997), pp. 267–299.

[38] Marc Niethammer, Yang Huang, and François-Xavier Vialard. "Geodesic Regression for Image Time-Series". In: *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2011.

[39] Sébastien Ourselin et al. "Block Matching: A General Framework to Improve Robustness of Rigid Registration of Medical Images". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000*. Springer. 2000, pp. 557–566.

[40] Xavier Pennec et al. "Riemannian Elasticity: A Statistical Regularization Framework for Non-Linear Registration". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*. Springer, 2005, pp. 943–950.

[41] Torsten Rohlfing et al. "Modeling Liver Motion and Deformation During the Respiratory Cycle Using Intensity-Based Nonrigid Registration of Gated MR Images". In: *Medical Physics* 31 (2004), p. 427.

[42] Daniel Rueckert et al. "Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images". In: *Medical Imaging, IEEE Transactions on* 18.8 (1999), pp. 712–721.

[43] N. Singh et al. "Genetic, Structural and Functional Imaging Biomarkers for Early Detection of Conversion from MCI to AD". In: *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2012, pp. 132–140.

[44] Martin Szegedi et al. "Four-Dimensional Tissue Deformation Reconstruction (4D TDR) Validation Using a Real Tissue Phantom". In: *Journal of Applied Clinical Medical Physics (JACMP)* 14.1 (Jan. 2013), pp. 115–132. ISSN: 15269914.

[45] D'Arcy Wentworth Thompson. *On Growth and Form*. Canto (Cambridge University Press). Cambridge University Press, 1917. ISBN: 9780521437769.

[46] Paul M Thompson and Arthur W Toga. "A Framework for Computational Anatomy". In: *Computing and Visualization in Science* 5.1 (2002), pp. 13–34.

[47] L Younes, F Arrate, and M I Miller. "Evolutions Equations in Computational Anatomy". In: *NeuroImage* 45.1 (2009), S40–S50.

CHAPTER 2

4D MAP IMAGE RECONSTRUCTION WITH INCOMPRESSIBLE MOTION MODEL

Reprinted from Medical Image Analysis 16 (2012) 1307–1316 with permission from Elsevier

Medical Image Analysis 16 (2012) 1307–1316

Contents lists available at SciVerse ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

4D CT image reconstruction with diffeomorphic motion model

Jacob Hinkle [a,*], Martin Szegedi [b], Brian Wang [b], Bill Salter [b], Sarang Joshi [a]

[a] Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, United States
[b] Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

ARTICLE INFO

Article history:
Received 22 July 2011
Received in revised form 18 May 2012
Accepted 31 May 2012
Available online 16 June 2012

Keywords:
Motion
Image reconstruction
Diffeomorphism
CT

ABSTRACT

Four-dimensional (4D) respiratory correlated computed tomography (RCCT) has been widely used for studying organ motion. Most current RCCT imaging algorithms use binning techniques that are susceptible to artifacts and challenge the quantitative analysis of organ motion. In this paper, we develop an algorithm for analyzing organ motion which uses the raw, time-stamped imaging data to reconstruct images while simultaneously estimating deformation in the subject's anatomy. This results in reduction of artifacts and facilitates a reduction in dose to the patient during scanning while providing equivalent or better image quality as compared to RCCT. The framework also incorporates fundamental physical properties of organ motion, such as the conservation of local tissue volume. We demonstrate that this approach is accurate and robust against noise and irregular breathing patterns. We present results for a simulated cone beam CT phantom, as well as a detailed real porcine liver phantom study demonstrating accuracy and robustness of the algorithm. An example of applying this algorithm to real patient image data is also presented.

Published by Elsevier B.V.

1. Introduction

Imaging of moving organs using conventional static reconstruction techniques is susceptible to motion artifacts for CT images. To alleviate these artifacts the raw data is often binned and individual 3D images are reconstructed which represent the anatomy in various configurations (Vedam et al., 2003; Rietzel et al., 2005; Brandner et al., 2006; Abdelnour et al., 2007). Inherent to the binning processes is that only a portion of the collected data is used for each reconstruction, compromising signal-to-noise ratio (SNR). This necessitates an increase in imaging dose in order to maintain sufficient image quality. Additionally, it has been well-established (Abdelnour et al., 2007) that binning processes are susceptible to artifacts due to insufficient data and irregular motion. To quantitatively analyze organ motion, it is now common practice to use deformable image registration in order to bring the binned images into correspondence and study the deformation fields (Foskey et al., 2005; Pevsner et al., 2006; Boldea et al., 2008). When the reconstructed images contain artifacts, image registration is challenged and the resulting motion estimate is compromised.

In most four dimensional (4D) imaging protocols, a motion surrogate signal is recorded during data acquisition for use in binning. This is often an external signal, such as a chest wall marker or circumference measurement for respiratory-correlated CT or EKG for cardiac imaging, but it may also be derived from the data, such as a measurement of diaphragm position from cone beam CT projections (Rit et al., 2008). Binning the data based on the amplitude of the surrogate signal is referred to as amplitude binning (Abdelnour et al., 2007). In practice, respiratory motion is nearly periodic, but exhibits hysteresis, meaning points in the subject's anatomy follow slightly different trajectories during inhale and exhale (Boldea et al., 2008; Langner and Keall, 2009). In the case that this difference is significant, the surrogate signal value does not accurately parametrize the motion. It is common to employ a binning method known as phase binning, in which the motion is assumed to be periodic. Periods are estimated using peak-detection on the surrogate signal, and the generated binned images correspond to distinct phase ranges (Abdelnour et al., 2007).

When respiratory motion is irregular, amplitude binning suffers from lack of data at some amplitudes. Additionally, in the presence of hysteresis, artifacts are introduced due to data mismatch. Phase binning is also challenged in the presence of irregular breathing. Artifacts appear when data have similar phases but very different amplitudes. The algorithm proposed in this paper avoids many of the inadequacies of binning processes by using all available data. The motion between projections is modeled so that data is compared in a consistent way, alleviating data mismatch artifacts caused by large bin widths. We also present a method for modeling amplitude-indexed motion in the presence of hysteresis.

Previous approaches to studying medical image motion fall into one of three groups: those that correct raw data on a frame-by-frame basis to try and decrease artifacts in a single 3D image reconstruction, those that use previously reconstructed 3D images

* Corresponding author. Tel.: +1 8015879660.
E-mail address: jacob@sci.utah.edu (J. Hinkle).

1361-8415/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.media.2012.05.013

such as binned images or breathhold images to track motion, and those that incorporate the motion model directly into the image reconstruction process without the use of a previously reconstructed 3D image. Members of the last group we refer to as four-dimensional (4D) image reconstruction algorithms. The framework presented in this work falls into the 4D reconstruction group. The framework employs a diffeomorphic motion model which can accomodate large deformation in compressible or incompressible tissue and is applicable to a variety of imaging modalities.

Among algorithms that aim to correct motion artifacts by correcting the raw in-plane data directly, Lu and Mackie (2002) fit a simple 2D scaling motion model to raw sinogram data and undo the distortion before applying filtered backprojection (FBP). Yu and Wang (2007) adopt a rigid 2D motion model for correcting head motion during acquisition to alleviate in-plane artifacts in fan beam CT. Ehrhardt et al. (2007) reconstruct 3D images at arbitrary breathing amplitudes by interpolating each slice from those collected at nearby amplitudes and then stacking them, using an in-plane optical flow model. Such simple in-plane motion models are not sufficient for imaging of the torso, where respiratory-induced motion causes highly non-linear deformation with the most significant component in the superior-inferior direction.

In order to address out-of-plane motion, algorithms in the second group focus on bringing conventionally reconstructed 3D images into correspondence with time-indexed imaging data. Zeng et al. (2007) and Li et al., 2007 presented B-spline-based methods that require an artifact-free reference image (such as a breath-hold image) in addition to a 4D RCCT fan beam or cone beam scan. Reyes et al. (2007) use two binned or gated MRI images (corresponding to inhale and exhale) to build a deformation model with which they correct positron emission tomography (PET) data then reconstruct a 3D PET image. These approaches address motion artifacts caused by slow scanning, but the acquisition of a breath-hold reference image causes unnecessary imaging dose and is impractical for many patients.

Instead of using breathhold images, Rit et al. (2008) use a retrospectively binned 4D cone beam scan to estimate organ motion using an optical flow model between 3D frames. Foskey et al. (2005) use a pairwise image registration framework in order to perform a registration between frames of a binned 4D CT dataset. They employ a diffeomorphic model similar to the one described in this paper, which ensures smoothness and invertibility while not penalizing large deformations. In both of these approaches, image registration is performed between each pair of adjacent binned images. Castillo et al. (2010) instead use a temporal spline-based trajectory model to estimate motion using a set binned 3D images. These algorithms all rely on an initial binning step, so motion artifacts due to the binning process result in false motion estimates.

Among algorithms which do not use a previously acquired base image, Mair et al. (2006) use an elastic optical flow motion model to estimate motion and images from gated cardiac CT data. In their algorithm, a different 3D image is estimated for every frame of a 4D image. Blume et al. (2010) estimate image and motion simultaneously for PET data. Their general approach is similar to the one presented in this paper. However, where they impose temporal smoothness on their estimated, time-varying deformation, the model we will present incorporates a full fluid flow model, capable of modeling large deformations and incompressible tissue in a consistent way.

Previous 4D reconstruction techniques have not incorporated sufficient modeling of incompressible tissue. Incompressibility has been studied, however, in the context of image registration. As early as 1991, Song and Leahy (1991) used an incompressible optical flow method for image registration. Rohlfing et al. (2003) use a spline-based model which penalizes tissue compression to perform incompressible image registration. Saddi et al. (2007) study incompressible fluid-based registration of liver CT. Their approach requires a solution of Poisson's equation via a multigrid method at each iteration. We build on this previous work by extending the diffeomorphic fluid flow model to include a hard incompressibility constraint, which we enforce through a Fourier approach similar to that of Stam (2001).

In this paper, we validate a 4D reconstruction method, extending the framework first introduced in Hinkle et al. (2009), which is based on a well-known diffeomorphic flow model (Christensen et al., 1996; Joshi, 1998; Beg et al., 2005; Foskey et al., 2005). The reconstruction of a base image and deformation velocity fields are phrased as a maximum a posteriori (MAP) estimation problem which is optimized using variational methods. The resulting problem can optionally be constrained to produce incompressible deformations either globally or in automatically determined regions. Our framework accommodates Gaussian or Poisson detector noise and can be optimized using whichever numerical scheme is appropriate. Scanner geometry is modeled abstractly as a linear projection, allowing our framework to be extended to CT, PET, and MRI modalities.

Our approach offers a number of advantages over previous work. It avoids binning algorithms altogether, allowing a significant reduction in CT dose while maintaining image SNR and reducing artifacts. It employs a realistic diffeomorphic tissue deformation model which is flexible enough to model compressible or incompressible tissue and large deformations. Furthermore, the framework is general and can be applied to many different four-dimensional imaging applications.

2. Methods and materials

Most conventional CT image reconstruction techniques estimate a static three-dimensional image, $I(\vec{x})$, which represents the patient's anatomy during data acquisition. Instead, our approach is to estimate a 4D time-indexed image, denoted by $I_t(\vec{x})$, capturing the subject's deforming anatomy. The following sections describe how this 4D image is estimated using data from an imaging device.

2.1. Data acquisition

Raw data from most imaging modalities can be described as a series of projections, $d_i \in \mathbb{C}^m,\ i = 1,\ldots,N$, representing measurement values taken from a set of m detector elements. A static image, $I(\vec{x})$, is modeled as a square-integrable function on a compact image domain $\Omega \subset \mathbb{R}^3$. The acquisition of projection data is modeled by a set of linear projection operators $P_i : L^2(\Omega) \to \mathbb{C}^m$, where $L^2(\Omega)$ denotes the space of square-integrable functions on Ω. Here we review the projection operators associated with the most common CT imaging modalities.

In fan beam CT, an X-ray source is mounted on a swinging gantry opposite a row of detectors. At each gantry angle, the detectors measure the attenuation of the X-ray intensity as photons pass through the subject. The image in this case represents the attenuation coefficient at each point in space, and the negative logarithms of the detector values depend linearly on the image (Prince and Links, 2006). This dependence is described by the Radon transform:

$$\{P_i I_{t_i}\}_j = \int I_{t_i}(s_j \cos\theta_{ij} + l \sin\theta_{ij}, s_j \sin\theta_{ij} - l\cos\theta_{ij}, z_i)dl, \tag{1}$$

where z_i is the slice location, s_j is the distance of the ray from the axis of rotation and θ_{ij} is the angle of the ray from X-ray source to detector j with respect to some reference direction.

Cone beam CT is similar to fan beam, however in this case a two-dimensional array of detectors is used. The projection operators

representing the 2D planar measurements of X-ray attenuation, $d_i(u,v)$ at a collection of angles, θ_i, are given by

$$\{P_i I_{t_i}\}_j = \int_0^\infty I_{t_i}(\vec{x}_{\theta_i} + l(\vec{w}_{\theta_i}(u_j, v_j) - \vec{x}_{\theta_i}))ldl \tag{2}$$

where $\vec{x}_{\theta_i}$ is the 3D X-ray source position and $\vec{w}_{\theta_i}(u_j, v_j)$ is the 3D position of the detector element j, which has planar coordinates (u_j, v_j) (Prince and Links, 2006; Feldkamp et al., 1984).

2.2. Noise model

Detector noise has been well-studied for most imaging modalities. Low-signal photon counting is subject to Poisson noise. Such is the case in low dose CT scans. Let $\mathcal{L}(d|I)$ denote the log-likelihood of observing the data d given the 4D image I. The form of the data log-likelihood depends on the noise model employed. For most imaging modalities with high enough SNR, the noise is approximately Gaussian. Under this assumption, the data log-likelihood is simply the sum of squared error:

$$\mathcal{L}(d|I) = -\frac{1}{\sigma^2}\sum_{i,j}|\{P_i I_{t_i}\}_j - d_{i,j}|^2, \tag{3}$$

where σ^2 is the noise variance.

For low-SNR imaging involving a photon-counting process, the noise is usually modeled by a Poisson distribution. The data log-likelihood under such an assumption takes the form

$$\mathcal{L}(d|I) = \sum_{i,j} d_{i,j}\ln[\{P_i I_{t_i}\}_j] - \{P_i I_{t_i}\}_j. \tag{4}$$

2.3. Motion model

Having modeled the detector geometry and noise, we could attempt to estimate an $I(t,\vec{x})$ which maximizes the data log-likelihood. Indeed this is the basis of many static reconstruction algorithms which estimate $I(\vec{x})$ in order to best fit the data. In the case of imaging moving anatomy, the additional temporal dimension of the image and the sparsity of data force us to look beyond a simple 4D maximum-likelihood reconstruction. Following Khan and Beg (Khan et al., 2008), we model the 4D image as a single 3D image $I_0 \in L^2(\Omega)$ undergoing a time-indexed deformation $g{:}[0,T]\times\Omega\to\Omega$. In this formulation $I(t,\vec{x})$ is written as $I_0\circ g^{-1}(t,\vec{x})$. The problem at hand is to then estimate both the base image and deformation which best fit the data. The estimated time-indexed deformation is meant to model the motion of the anatomy during image acquisition. As organs are not expected to tear apart or change their topology during physiological motion, we model the time-indexed deformation as a flow along smooth velocity fields $w(t,\vec{x})$, which are defined as

$$w(t,\vec{x}) = \frac{d}{dt}g(t,\vec{x}). \tag{5}$$

If the velocity fields are spatially smooth, then the resulting deformation is guaranteed to be a diffeomorphism, a bijective smoothly differentiable mapping (Arnol'd, 1989). This ensures that embedded structures transform in a feasible manner. Note that relative motion of organs in the form of sliding is not permitted in this model. However, other forms of regularization that accomodate sliding can be applied within this 4D reconstruction framework, for instance the L^1 curl penalty approach recently developed by Ruan et al. (2009).

Given a set of smooth velocity fields, the deformation can be recovered by integrating:

$$g(t,\vec{x}) = \vec{x} + \int_0^t w(\tau, g(\tau,\vec{x}))d\tau. \tag{6}$$

In order to enforce smoothness, the velocity fields are estimated subject to a regularization term that has the form of a squared Sobolev norm,

$$\|w\|_V^2 = \langle w, w\rangle_V = \int_0^T \int_\Omega \|Lw(t,\vec{x})\|^2_{\mathbb{R}^3} d\vec{x}dt, \tag{7}$$

where L is a differential operator chosen to reflect physical tissue properties. The interpretation of Eq. (7) as the negative log-probability of a formal prior deserves some explanation. Following Kuo (1975), we place a Gaussian random field prior with covariance $(L^\dagger L)^{-1}$ on the Sobolev space associated to the differential operator L. By the Sobolev embedding theorem, this Sobolev space is embedded in a Banach space of continuous vector fields. The continuity properties of elements in the Banach space are determined by the choice of Sobolev space, which in turn is determined by the order of the differential operator L. In our implementation, $Lw = -\alpha\nabla^2 w - \beta\nabla\nabla\cdot w + \gamma w$ for scalar parameters α, β, γ, following Christensen et al. (1996), Beg et al. (2005), Davis (2008), guaranteeing that solutions to the minimum norm problem will be continuous. The α and β parameters penalize the Laplacian and divergence, respectively, of the velocity field, while the γ term is regularizing in the sense that it ensures that L is a positive definite operator (Davis, 2008). Note that higher-order differential operators could be used to ensure higher orders of differentiability in the resulting velocity fields. For a more detailed discussion of the choice of differential operator see Amit et al. (1991), Dupuis et al. (1998), Trouvé (1995). The reader is referred to Dupuis et al. (1998), Budhiraja et al. (2010) for a detailed discussion of the minimum norm problem presented here as a Bayesian estimation problem.

As mentioned earlier, in most 4D imaging protocols, a surrogate signal, $a(t)$, correlated with the deforming patient anatomy is simultaneously recorded. If the signal, $a(t)$, is a faithful surrogate of the internal anatomical configuration, then the deformation can be parametrized by $a(t)$ instead of t. We introduce the deformation h, parametrized by the surrogate signal, defined by

$$h(a(t),\vec{x}) = g(t,\vec{x}) \tag{8}$$

Returning to the definition of $w(t,\vec{x})$ the change of variables is given by:

$$w(t,\vec{x}) = \frac{d}{dt}h(a(t),\vec{x}) = v(a(t), h(a(t),\vec{x}))\frac{da}{dt}, \tag{9}$$

where $v(a,h(a,\vec{x})) = \frac{\partial}{\partial a}h(a,\vec{x})$ is a velocity field with respect to changes in surrogate signal instead of time. The deformation to any amplitude is then obtained by the integration of Eq. (9),

$$h(a,\vec{x}) = \vec{x} + \int_0^a v(a', h(a',\vec{x}))da'. \tag{10}$$

As discussed previously, patient anatomy cannot usually be fully parametrized by the surrogate signal amplitude alone (Boldea et al., 2008). In fact, it has been demonstrated that signal amplitude, along with the time derivative of the signal give a more consistent parametrization of respiratory motion (Langner and Keall, 2009). In the presence of significant hysteresis, we model the 4D image as a base image at one end of the hysteresis loop, along with two sets of vector fields: one representing inhale motion and the other representing exhale. The raw data is separated based on whether the time derivative of the surrogate was positive (indicating inhale) or negative (indicating exhale). The resulting estimate provides two different deformations $h_{in}(a,\vec{x})$ and $h_{ex}(a,\vec{x})$ representing inhale and exhale motion, respectively.

2.4. Posterior log-probability and MAP estimation

Following the MAP paradigm, the data log-likelihood and motion prior are combined to give the posterior log-probability

$$\mathcal{L}(I_0, v|d_i) = -\|v\|_V^2 - \frac{1}{2\sigma^2}\sum_{ij}|\{P_i I_0 \circ h_{a_i}^{-1}\}_j - d_{ij}|^2. \quad (11)$$

The 4D image reconstruction problem is to estimate the image and velocity fields that maximize the posterior,

$$(\widehat{I}_0, \hat{v}) = \underset{I_0, v}{\operatorname{argmax}}\, \mathcal{L}(I_0, v|d_i). \quad (12)$$

A MAP estimate is obtained via an alternating iterative algorithm which at each iteration updates the estimate of the deformation in a gradient descent step then updates the image. For Poisson noise, the posterior can be maximized via an expectation–maximization (EM) algorithm instead of the gradient descent version. In such case, EM iterations may be used for the image update steps, while interleaving steepest descent iterations on the velocity fields. In such a case, care should be taken to control the stability of the algorithm by limiting the number of image update iterations, in order to avoid blow up due to the ill-conditioned nature of the EM algorithm in the absence of an image prior. We have found that for Poisson noise, stable convergence is obtained by first performing a number of gradient descent iterations on both base image and velocity fields before applying EM. Another strategy is to place a prior on the base image as well as the velocity fields. The image prior can take the form a regularization term such as the total variation or Good's roughness functionals. See, for instance, (Vardi and Lee, 1993) and (Miller and Younes, 2001) for a more detailed explanation of these stability issues.

2.5. Implementation details

The continuous amplitude-indexed velocity field is discretized by a set of equally-spaced surrogate signal values a_k with the associated velocities v_k and spacing Δa. This discretization is independent of the signal values at which data is acquired. The inverse deformation from a_{k-1} to a_k is approximated by the backward Euler integration of Eq. (10),

$$h_{a_{k-1}}(h_{a_k}^{-1}(\vec{x})) \approx x - \Delta a\, v_{k-1}(x) \quad (13)$$

which can be applied k many times to approximate the inverse deformation $h_{a_k}^{-1}$. The deformation for an amplitude a between a_{k-1} and a_k is linearly interpolated as

$$h_{a_{k-1}}\left(h_a^{-1}(\vec{x})\right) \approx x - (a - a_{k-1})\, v_{k-1}(x). \quad (14)$$

Higher-order integration schemes such as Runge–Kutta may also be used in place of the simpler Euler method.

The amplitude steps, Δa, must be chosen by the user. A very small choice of Δa better ensures that the deformation will be smoothly varying and invertible. However, as Δa decreases, more velocity fields must be estimated, which increases computational complexity.

2.5.1. Optimization of velocity fields

For any amplitude a, the Sobolev first variation of Eq. (11) with respect to v_a under the inner product determined by L is given by

$$\delta_{v_a}\mathcal{L} = -2v_a + \frac{1}{\sigma^2}K\sum_{i,a<a_i}|Dh_{a,a_i}|P_i^\dagger\left\{P_i\left\{I_0 \circ h_{a_i}^{-1}\right\} - d_i\right\} \circ h_{a,a_i}\nabla I_a, \quad (15)$$

where $h_{a,a_i} = h_{a_i} \circ h_a^{-1}$ is the deformation from a to a_i, $K = (L^\dagger L)^{-1}$ is the smoothing operator associated with L, and $P_i^\dagger$ is the adjoint of the projection operator, which acts by backprojecting the data discrepancy into the 3D volume. Note that the Sobolev variation is distinguished from the more familiar L^2 variation by the presence of the smoothing operator K. Though either gradient could be used in a gradient descent scheme, the high frequency suppression of the Sobolev variation leads to a more stable numerical scheme, as discussed in Beg (2003). Using the amplitude discretization in Eqs. 13 and 14, the variation with respect to the kth velocity field is

$$\begin{aligned}\delta_{v_k}\mathcal{L} = -2v_k &+ \frac{1}{\sigma^2}K\sum_{i,a_k<a_i\leqslant a_{k+1}}\frac{a_i - a_k}{\Delta a}|Dh_{a_k,a_i}(x)|B_i(x)\nabla I_{k-1} \\ &\times (x - (a_i - a_k)v_k(x)) + \frac{1}{\sigma^2}K\sum_{i,a_{k+1}<a_i}|Dh_{a_k,a_i}(x)|B_i(h_{a_{k+1},a_i}(x)) \\ &\times \nabla I_{k-1}(x - \Delta a\, v_k(x)) \end{aligned} \quad (16)$$

where

$$B_i = P_i^\dagger\left\{P_i\left\{I_0 \circ h_{a_i}^{-1}\right\} - d_i\right\} \quad (17)$$

Note that for each piece of data, d_i, we compare the pushed-forward image to that data then backproject into the volume to compute B_i. We then pull the result back to a_k, which involves multiplying by the Jacobian determinant as we interpolate by h_{a_k,a_i}. The resulting vector fields are then smoothed by the operator K and summed.

Following the approach of Beg et al. (2005), efficient computation of K is implemented in the Fourier domain, requiring only Fourier transforms of v_k followed by a matrix multiplication at each point, at each iteration of the algorithm.

2.5.2. Optimization of base image

The first variation of Eq. (11) with respect to I_0 for Gaussian noise is

$$\delta_{I_0}\mathcal{L}(I_0, v|d_i) = -\frac{1}{\sigma^2}\sum_i |Dh_{a_i}|P_i^\dagger\left(P_i\left\{I_0 \circ h_{a_i}^{-1}\right\} - d_i\right) \circ h_{a_i}, \quad (18)$$

whereas for Poisson noise the following gradient may be used in a gradient descent:

$$\delta_{I_0}\mathcal{L}(I_0, v|d_i) = \sum_i |Dh_{a_i}|P_i^\dagger\left(\frac{d_i}{P_i\left\{I_0 \circ h_{a_i}^{-1}\right\}} - 1\right) \circ h_{a_i}. \quad (19)$$

During the gradient descent algorithm, the velocity fields and base image are updated by first calculating these variations then multiplying by a fixed step size and adding. In the presence of Poisson noise, care should be taken that the image I_0 is constrained to be everywhere non-negative if implementing gradient descent.

As discussed previously, the image can alternatively be estimated using an ML-EM algorithm, characterized by multiplicative updates of the form

$$I_0^{k+1} = I_0^k \frac{\sum_i |Dh_{a_i}|P_i^\dagger\left\{\frac{d_i}{P_i\left\{I_0^k \circ h_{a_i}^{-1}\right\}}\right\} \circ h_{a_i}}{\sum_i |Dh_{a_i}|P_i^\dagger\{1\} \circ h_{a_i}}. \quad (20)$$

As soon as the velocity field is updated, the image estimate must also be updated. The change of image estimate in turn alters the velocity gradients, leading to a joint estimation algorithm in which, at each iteration, the velocity fields are updated and the image recalculated.

2.6. Incompressibility constraint

As discussed previously, it is sometimes useful to enforce further tissue constraints. When modeling organs such as the liver, which is essentially incompressible during normal activity,

unrealistic deformations may be easily recognized if they represent local compression or expansion. Estimation of these types of unrealistic deformations may be avoided by constraining the deformations to be incompressible. Deformations defined as a flow along smoothly-varying vector fields as described in Eq. (8) have been well studied (Arnol'd, 1989). In particular, if the divergence of the velocity field is zero, the resulting deformation is guaranteed to preserve volume locally and have unit Jacobian determinant.

The Helmholtz-Hodge decomposition allows us to implement the incompressibility constraint by projecting the unconstrained velocity fields onto the space of divergence-free vector fields at each iteration of the algorithm (Cantarella et al., 2002). In order to efficiently implement the Helmholtz-Hodge decomposition of a time-varying velocity field, we use the discrete divergence operator as it operates in the Fourier domain. We write the discrete Fourier transform of a central difference approximation to the derivative of a 1D function f as

$$\begin{aligned}\mathrm{DFT}\{\delta_x f\}(\omega) &= \mathrm{DFT}\left\{\frac{f(x+\Delta x)-f(x-\Delta x)}{2\Delta x}\right\}(\omega)\\ &= \frac{i}{2\Delta x}\sin\omega \mathrm{DFT}\{f\}(\omega),\end{aligned} \tag{21}$$

where Δx is the grid spacing in the x direction. Applying this formula to each component of a vector field v, we see that the divergence of v takes the following form in the Fourier domain:

$$\mathrm{DFT}\{\mathrm{div}\, v\}\ (\vec{\omega}) = W(\vec{\omega})\cdot \mathrm{DFT}\{v\}(\vec{\omega}), \tag{22}$$

where

$$W(\vec{\omega}) = \frac{i}{2}\begin{pmatrix} \frac{1}{\Delta x}\sin\frac{\omega_x}{N_x} \\ \frac{1}{\Delta y}\sin\frac{\omega_y}{N_y} \\ \frac{1}{\Delta z}\sin\frac{\omega_z}{N_z}\end{pmatrix} \tag{23}$$

and the dot indicates inner product of complex vectors in $\mathbb{C}^3$. This allows us to remove the divergent component easily in Fourier space via the projection

$$\mathrm{DFT}\{v\} \mapsto \mathrm{DFT}\{v\} - \left(\frac{W\cdot \mathrm{DFT}\{v\}}{\|W\|^2}\right)W. \tag{24}$$

This projection corresponds to the L^2 inner product on vector fields. A projection corresponding to the Sobolev inner product induced by L may also be used, but for simplicity we have employed the L^2 projection only. Since the operator K is implemented in the Fourier domain, the Fourier transform of v is already computed and there is little computational overhead in performing this projection at each iteration of the algorithm described in Section 2.4.

2.6.1. Local incompressibility constraint

The Fourier method efficiently enforces incompressibility over the entire image domain. However, when the field of view includes both compressible and incompressible tissues, this leads to unrealistic deformation estimates in some regions. Given a mask M: $\Omega \to [0,1]$, a compressible velocity field update dv_c and its incompressible projection dv_i, the mask can be used to combine the two velocity updates via the convex combination

$$v(x) \mapsto v(x) + M(x)dv_c(x) + (1-M(x))dv_i(x). \tag{25}$$

This simple linear combination of velocity field updates constitutes a stable numerical scheme, while the mask M allows the user to model image regions as compressible or incompressible. Alternatively, the mask can be automatically generated using a blurred threshold of the deforming image. As a result, individual organs can be modeled as either compressible or incompressible without requiring user intervention and while requiring relatively little computation compared to the globally incompressible algorithm.

2.7. 4D reconstruction from slice data

If single-slice acquisition time is fast compared to organ motion, then individual slices are reconstructed with few artifacts using filtered backprojection. This assumption holds reasonably well in the case of 4D RCCT. In this case the 4D image is estimated by fitting a deforming 3D image to the reconstructed 2D slices $S_i(x,y)$ instead of directly to the raw sinogram data. As previously discussed, the sinogram will exhibit spatially independent Poisson or approximately Gaussian noise. Filtered backprojection reconstruction complicates the noise properties in the resulting slice. However, Wilson and Tsui (1993) have shown that the reconstructed noise is approximately Gaussian with variance independent of pixel intensity, justifying the use of a Gaussian noise model for slice data. Under this slow motion assumption, the velocity field variation becomes

$$\delta_{v_a}\mathcal{L} = -2v_a - \frac{1}{\sigma^2}K\sum_{\substack{(h_{a,a_i})_z = z_i \\ a_i > a}} |Dh_{a,a_i}|\left(I_0\circ h_{a_i}^{-1}(x,y,z) - S_i(x,y)\right)\circ h_{a,a_i}\nabla I_a. \tag{26}$$

Assuming the incompressibility constraint is enforced, for a given set of velocity fields the base image that maximizes Eq. (18) can be computed in closed form, and corresponds to arithmetic mean of the deformed data

$$\widehat{I}_0(\vec{x}) = \frac{1}{N}\sum_{i,h(a_i,\vec{x})_z = z_i} S_i \circ h_{a_i}(\vec{x}). \tag{27}$$

This base image estimate solves the Euler–Lagrange equation for Eq. (11). If the incompressibility constraint is not enforced, the mean in Eq. (27) becomes a weighted mean with weights given by the Jacobian determinants of the deformations h_{a_i}.

3. Results

3.1. Fan beam CT phantom study

In order to validate the accuracy of the 4D reconstruction algorithm, a phantom study was performed using the CIRS anthropomorphic thorax phantom (CIRS Inc., Norfolk, VA) and a GE Lightspeed RT16 CT scanner (GE Health Care, Waukesha, WI). The phantom includes a simulated chest cavity with a 2 cm spherical object representing a tumor that is capable of moving in three dimensions. A chest marker is also included in the phantom which moves in a pattern synchronized to the tumor. The marker is tracked by a Real-time Position Management (RPM) system (Varian Oncology Systems, Palo Alto, CA) to generate respiratory signals for use in binning. A description of the GE phase-binned 4D-CT acquisition and processing has been previously published by Pan et al. (2004). For the scans in this study, the phantom was driven to simulate a breathing trace collected from a real patient.

Fig. 1 shows the stationary spherical CIRS lung tumor phantom, imaged with helical CT. In order to demonstrate robustness against noise, an initial scan was taken with an X-ray tube current of 250 mA, then repeated with a tube current of 25 mA. The 4D phase-binned image set generated by the GE Advance Workstation is shown in the top row of Fig. 2. Notice the binning artifacts, including mismatched slices in the phase-binned images. Also shown in the bottom row of Fig. 2 are images from an amplitude-binned dataset at peak-inhale, mid-range amplitude, and peak-exhale. These images exhibit mismatch artifacts because of the wide bin widths at peak-inhale and peak-exhale, which are necessary because of lack of data at those amplitudes.

Because we did not have access to the raw projection data, we applied the slow-motion assumption described previously when using the slice data along with the recorded RPM trace in the 4D

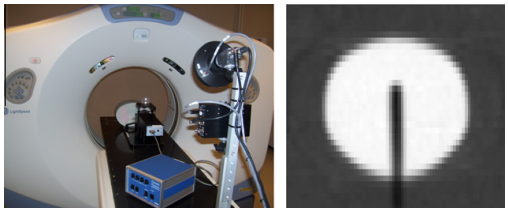


Fig. 1. CIRS phantom during scan setup with Varian RPM camera and GE Lightspeed RT scanner (left) and a helical CT scan of the stationary phantom (right).

reconstruction algorithm. Shown in Fig. 3 are the images reconstructed using our algorithm and the same raw data as the phase- and amplitude-binned images in Fig. 2. The reconstructed images do not have any artifacts such as those seen in the binned images. Notice also the increase in SNR in the 4D reconstructed images. 4D images reconstructed using our algorithm from the 25 mA data have higher signal-to-noise ratio (SNR = 76.5) than binned images reconstructed using the 250 mA data (SNR = 53.9). This shows that by using all of the data we are able to provide a higher SNR than current binning methods, while reducing the diagnostic dose to 10%. The similarity in images between the two 4D reconstructions shows the robustness of the image estimation to increasing noise.

To validate the estimated deformation model, a single point at the center of the phantom indicated by the cross-hair in Fig. 4 was tracked by integrating the estimated velocity fields according to Eq. (10). The physical construction of the phantom dictates that the superior-inferior displacement is linearly correlated to the RPM signal. Shown in Fig. 4 is a plot of the estimated displacements vs. the RPM signal. Notice the excellent linear correlation ($r = 0.9988$) between them, validating that the deformation estimation process leads to accurate point trajectory estimates.

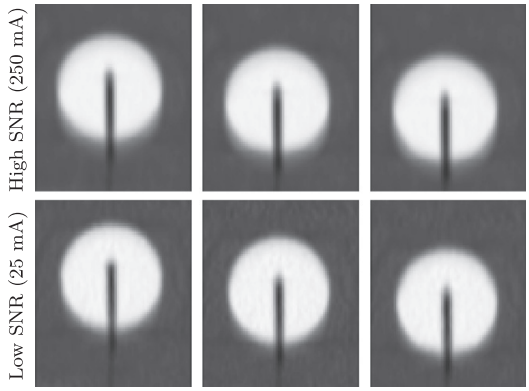


Fig. 3. Images of the phantom reconstructed using our 4D MAP image reconstruction algorithm at end-inhale, mid-range, and end-exhale amplitudes. The top row shows the 4D reconstruction of the high SNR data, while the bottom row shows that of the low SNR data. The 25 mA data are reconstructed with SNR similar to that of the binned 250 mA images.

3.2. Simulated cone beam phantom

A cone beam CT scan was simulated using an analytical 3D reconstruction phantom resembling the classical 2D Shepp-Logan phantom. Low SNR Poisson noise was generated in order to simulate the noise characteristics of a real scan. The data, consisting of 360 cone beam projections, each of size 64×64 pixels, were used to perform an iterative maximum likelihood (ML) EM 3D reconstruction (Vardi and Lee, 1993) of a volume of size $80 \times 80 \times 80$ voxels. The phantom was programmed to deform while data was collected. The velocity of the deformation is depicted in the top right of Fig. 5. The non-rigid deformation was designed to challenge the reconstruction process. A midsagittal slice of a static reconstruction using the motion and noise-corrupted data is shown in the top center of Fig. 5, while the bottom row shows the simulated projection data. The static 3D reconstruction exhibits extreme blurring artifacts of the internal structures as well as at the edge of the phantom.

The 4D MAP algorithm was run using the deforming phantom data set. Fig. 6 shows midsagittal slices of the estimated deforming image. Note the dramatic increase in image quality over the static reconstruction. As mentioned previously, since the 4D MAP algorithm uses all available data, the noise properties are similar to that of the ML reconstruction, while avoiding the motion artifacts that result from inherently 3-dimensional methods.

3.3. Porcine liver phantom

In order to test the hysteresis estimation method, another phantom study was performed. In order to create a quantitative validation approach which avoided inter-observer influence, we used a previously described motion phantom containing a porcine liver lobe with markers embedded for voxel-to-voxel accuracy testing of our 4D reconstruction algorithm (Szegedi et al., 2010). The liver is immersed in a container filled with Krebs Henseleit (KH) fluid (Krebs and Henseleit, 1932). The phantom moves a diaphragm surrogate, exerting force onto porcine liver tissue placed between a fixed support and a moving support. Implanted fiducials show that the phantom reproduces liver motion that is equivalent to respiratory-driven human liver motion measured in patients undergoing radiation treatment (Szegedi et al., 2009). In particular, the

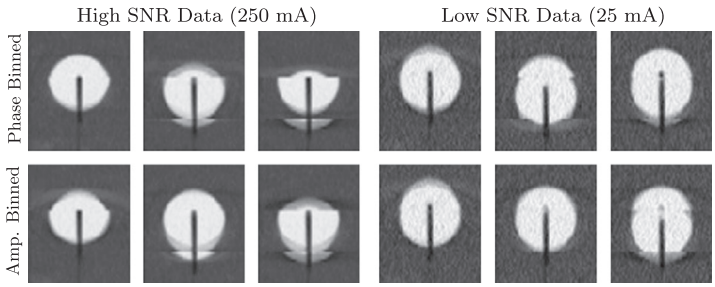


Fig. 2. Binned images of the moving CIRS phantom. The top row shows phase binned images at three different phases for the high SNR data (left) and the low SNR (10% tube current) data (right). The bottom row shows amplitude binned images at end-inhale, mid-range, and end-exhale amplitudes for both the high and low SNR data. Both phase and amplitude binning result in significant motion artifacts, even though the phantom was programmed to exhibit a perfect amplitude-motion correlation.

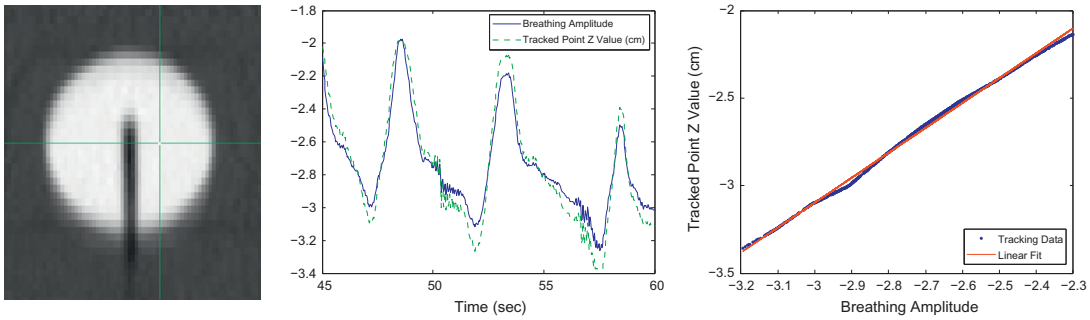


Fig. 4. Tracked point (left) with RPM signal and superior-inferior (z) coordinate (center) and plot of tracked point z coordinate vs. RPM signal (right) showing strong linear correlation ($r = 0.9988$).

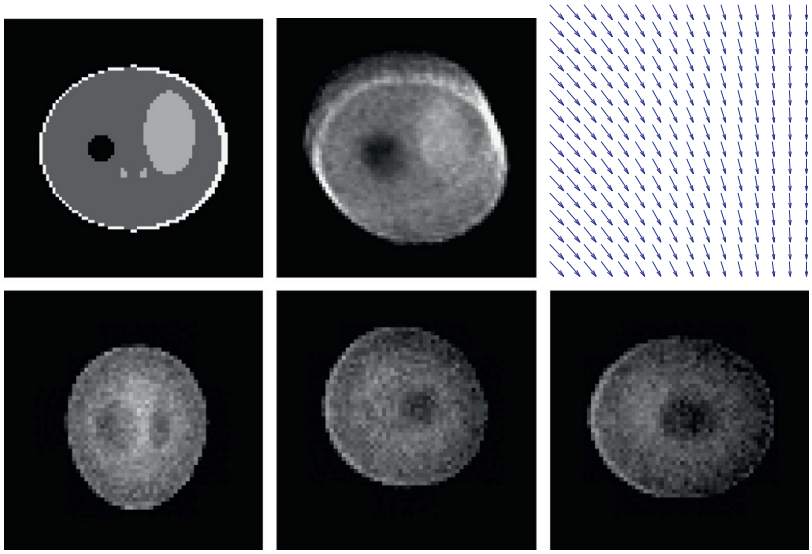


Fig. 5. Midsagittal slice of the simulated cone beam phantom (top left), static reconstruction of deforming phantom (top center), velocity field used to generate deformation (top right), and Poisson noise-corrupted cone beam projections of deforming phantom at 0, 45, and 90 degrees (bottom row). The static reconstruction shows severe blurring artifacts due to motion.

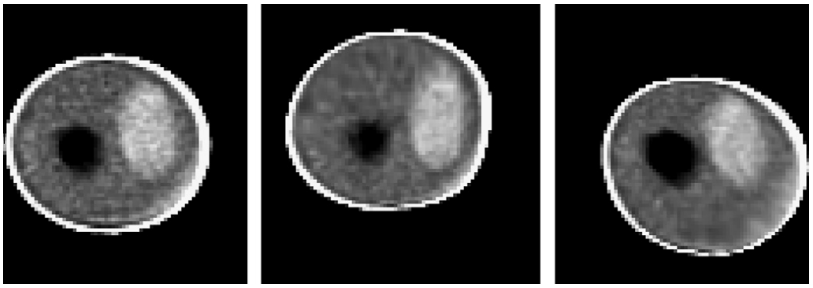


Fig. 6. Mid-sagittal slices of images reconstructed with our 4D reconstruction algorithm. The left image is the base image, while the other images are the interpolated base image at either end of the respiratory cycle. The reconstructed images show a crisp boundary and faithfully represent the deforming phantom. Note that the interpolated images appear slightly less noisy than the base image as a result of smoothing caused by linear interpolation.

phantom's liver motion shows hysteresis, in that the inhale trajectory does not follow exactly the exhale trajectory.

The 4DCT protocol used was our standard clinical stereotactic body radiation therapy (SBRT) protocol (1.25 mm slice thickness).

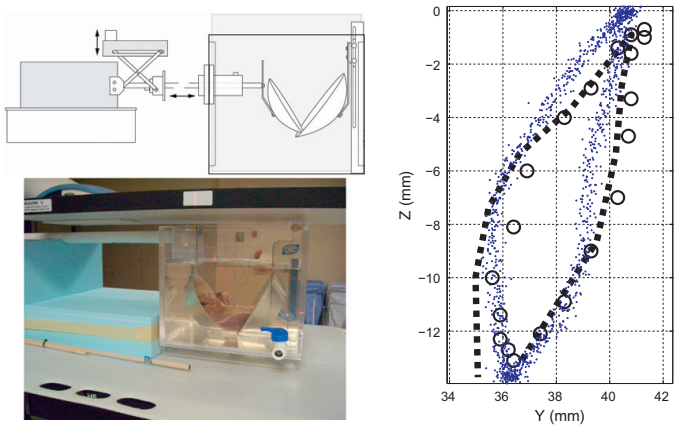


Fig. 7. Porcine liver phantom schematic (top left) and photo showing phantom in place inside the KH solution bath (bottom left). On the right are shown fiducial trajectories, measured manually (circles), EMT (dots), 4D MAP (dashed line). Notice the good agreement between the estimated 4D MAP trajectory and the ground truth (EMT) and manually delineated trajectories.

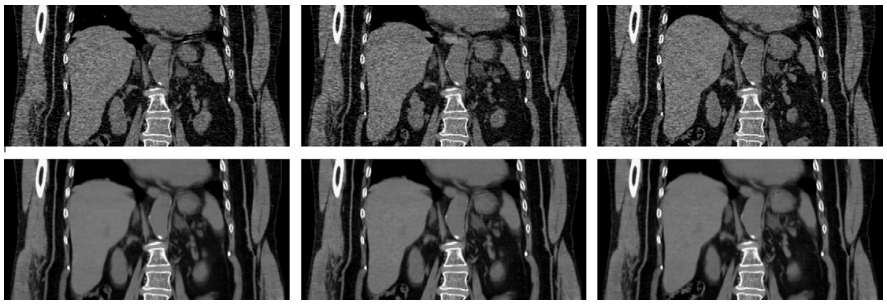


Fig. 8. Patient phase binned images (top) along with images reconstructed using the 4D MAP reconstruction algorithm (bottom) at peak-exhale, mid-range, and peak-inhale. The 4D reconstruction provides greatly increased SNR and consistency between images.

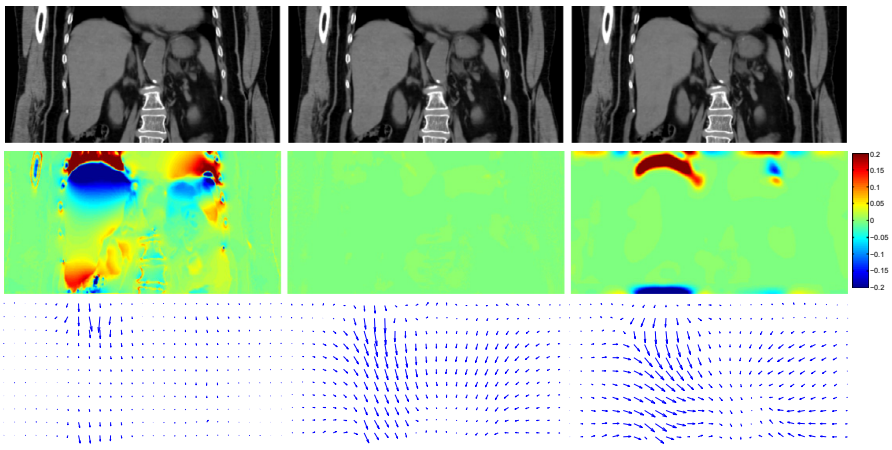


Fig. 9. Images reconstructed using our algorithm (top), log-Jacobian determinant images (center), and velocity fields (bottom), for compressible flow reconstruction (left), with incompressibility constraint (center), and with the mask-based approach (right). Negative log-Jacobian values indicate local compression, while positive values indicate expansion. The compressible algorithm predicts volume change at the top and bottom of the liver, but very little motion near the center. The incompressible constraint results in a nearly identical image estimate but predicts that the whole liver moves during breathing. The mask-based algorithm provides a similar motion estimate for the liver, while allowing for volume change in the lungs and intestines.

However, in order to achieve higher pixel resolution and maximize the number of images per slice position, we used a smaller field of view. This resulted in a reduced potential for error within the GE 4DCT phase-sampling algorithm.

Before acquiring the 4DCT, the phantom was prepared with a porcine liver containing electromagnetic tracking (EMT) transponders. The RPM surrogate marker box was placed on the chest-motion platform of the phantom during CT acquisition. The phantom was programmed to exhibit 6 s periodic sinusoidal motion. The 4D image data was processed using GE AW Sim MD software (version 7.6) to create 20 phase-binned images.

The 3D positions of the markers were tracked manually on each of the 20 phase-binned images. The CINE data along with the RPM trace was also used to generate a 4D MAP reconstructed image, using the hysteresis correction described in Section 2.5. The resulting motion estimate was used to reconstruct the marker trajectory.

In order to confirm that the 4DCT measurement of marker motion was accurate and that it did not suffer from phase-binning errors, we acquired a second, alternative measurement of marker motion using a wireless 4D EMT localizing and tracking system (Calypso Medical Technologies Inc., Seattle, WA, USA). The phantom was moved to an EMT-enabled treatment vault after the 4DCT acquisition where it was EMT tracked for multiple motion cycles. The EMT-generated data, which is of higher temporal resolution (10 Hz sampling frequency) than the 4DCT, was compared to the 4DCT-measured fiducial coordinates motion pattern and to the 4D MAP reconstruction predicted motion pattern.

Fig. 7 shows a projection of the measured and computed 3D trajectories to the *YZ* plane (there was less than 3 mm transponder motion in the *X* direction). Notice that the EMT and manual measurements agree quite well. Also notice that the hysteresis corrected algorithm estimates a trajectory (dashed curve) that very accurately matches the ground truth and is within 1 mm of both the manual measurements and EMT points.

3.4. Patient study

The 4D reconstruction algorithm was also applied retrospectively to data collected from a patient undergoing hypo-fractionated radiation therapy treatment of the liver at the Huntsman Cancer Institute at the University of Utah. The data consisted of 2991 slices, each of size 512×512 pixels, constituting 176 slice locations. The 4D reconstruction was performed using an amplitude discretization which divided the amplitude range equally into 10 parts. The multithreaded gradient descent algorithm converged in under 400 iterations and took roughly 23 h to run on a 32 core machine. However, this run time can be improved in a multiscale algorithm which processes progressively more detailed data while upscaling the base image and velocity field estimates. Through such an approach we have improved the running time to six or less hours. If further performance improvements are needed, it is possible to compute the algorithm on a region of interest within the image volume.

A comparison between phase binning and the 4D reconstruction is shown in Fig. 8. In addition to improving SNR, slice mismatch artifacts are absent in the 4D reconstructed image. The 4D reconstruction algorithm was run with and without the incompressibility constraint and also with the automatically thresholded mask-based incompressibility constraint described in Section 2.6.1. Fig. 9 shows an analysis of the incompressibility projection. The top row of the figure shows the base image deformed to the end of the amplitude range. The reconstructed images are extremely similar, while the Jacobian maps shown in the second row are quite different. In particular, it is seen in the third row that without the incompressibility constraint, the algorithm estimates compression and expansion of the top and bottom of the liver, while the incompressible reconstruction shows no local expansion or contraction. Given that liver is a blood-filled organ, physiologically it does not undergo any appreciable local changes in volume due to breathing. This exemplifies the necessity of incorporating incompressibility into the reconstruction process. Notice that the masked incompressible constrained motion indicates expansion of the lungs while still predicting that the entire liver moves during breathing.

4. Conclusion

A 4D reconstruction method was shown to produce artifact-free images with increased SNR over current binning methods. The increase in SNR is important in the case of CT, as it enables dose reduction to the patient during scanning. This reduction in dose enables more 4D scans to be acquired for patients undergoing radiation therapy. The reconstruction was also shown in phantom testing to provide an accurate estimate of the anatomical deformation taking place.

It is important to note that in the current implementation, since the deformations are modeled as diffeomorphic laminar fluid flows, there is no accommodation for organs sliding against one another. However, the general framework accommodates this by allowing a different motion prior which does not employ a homogeneous smoothness penalty, but instead allows discontinuities at organ boundaries. We demonstrated an extension to the algorithm that allows a more accurate motion estimate in the presence of significant hysteresis. This extension was shown to very accurately estimate a hysteresis loop observed in a porcine liver phantom.

The intent of this work is to introduce a framework by which 4D data is reconstructed using a motion surrogate signal. However, the framework is flexible and could be extended for other applications. For instance, if an artifact-free base image is available, this can be directly used to obtain an estimated deformation which best fits the data. Also, if enough data is available, the motion surrogate could be ignored and time-indexed deformations estimated.

References

Abdelnour, A.F., Nehmeh, S.A., Pan, T., Humm, J.L., Vernon, P., Schöder, H., Rosenzweig, K.E., Mageras, G.S., Yorke, E., Larson, S.M., Erdi, Y.E., 2007. Phase and amplitude binning for 4D-CT imaging. Phys. Med. Biol. 52, 3515–3529.

Amit, Y., Grenander, U., Piccioni, M., 1991. Structural image restoration through deformable templates. J. Am. Stat. Assn. 86, 376–387.

Arnol'd, V.I., 1989. Mathematical Methods of Classical Mechanics, second ed. Springer.

Beg, M.F., 2003. Variational and Computational Methods for Flows of Diffeomorphisms in Image Matching and Growth in Computational Anatomy. Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comp. Vis. 61, 139–157.

Blume, M., Martinez-Moller, A., Keil, A., Navab, N., Rafecas, M., 2010. Joint reconstruction of image and motion in gated positron emission tomography. IEEE Trans. Med. Imag. 29, 1892–1906.

Boldea, V., Sharp, G.C., Jiang, S.B., Sarrut, D., 2008. 4D-CT lung motion estimation with deformable registration: quantification of motion nonlinearity and hysteresis. Med. Phys. 35, 1008–1018.

Brandner, E.D., Wu, A., Chen, H., Heron, D., Kalnicki, S., Komanduri, K., Gerszten, K., Burton, S., Ahmed, I., Shou, Z., 2006. Abdominal organ motion measured using 4D CT. Int. J. Radiat. Oncol. Biol. Phys. 65, 554–560.

Budhiraja, A., Dupuis, P., Maroulas, V., 2010. Large deviations for stochastic flows of diffeomorphisms. Bernoulli 16, 234–257.

Cantarella, J., DeTurck, D., Gluck, H., 2002. Vector calculus and the topology of domains in 3-space. Am. Math. Monthly 109, 409–442.

Castillo, E., Castillo, R., Martinez, J., Shenoy, M., Guerrero, T., 2010. Four-dimensional deformable image registration using trajectory modeling. Phys. Med. Biol. 55, 305.

Christensen, G.E., Rabbitt, R.D., Miller, M.I., 1996. Deformable templates using large deformation kinematics. IEEE Trans. Imag. Proc. 5, 1435–1447.

Davis, B.C., 2008. Medical Image Analysis via Frèchet Means of Diffeomorphisms. Ph.D. thesis, University of North Carolina, Chapel Hill.

Dupuis, P., Grenander, U., Miller, M.I., 1998. Variational problems on flows of diffeomorphisms for image matching. Quart. J. Appl. Math. 56, 587–600.

Ehrhardt, J., Werner, R., Säring, D., Frenzel, T., Lu, W., Low, D., Handels, H., 2007. An optical flow based method for improved reconstruction of 4D CT data sets acquired during free breathing. Med. Phys. 34, 711–721.
Feldkamp, L.A., Davis, L.C., Kress, J.W., 1984. Practical cone-beam algorithm. J. Opt. Soc. Am. A 1, 612–619.
Foskey, M., Davis, B., Goyal, L., Chang, S., Chaney, E., Strehl, N., Tomei, S., Rosenman, J., Joshi, S., 2005. Large deformation three-dimensional image registration in image-guided radiation therapy. Phys. Med. Biol. 50, 5869–5892.
Hinkle, J., Fletcher, P.T., Wang, B., Salter, B., Joshi, S., 2009. 4D MAP image reconstruction incorporating organ motion. In: IPMI 2009: Proceedings of Information Processing in Medical Imaging, pp. 676–687.
Joshi, S.C., 1998. Large Deformation Diffeomorphisms and Gaussian Random Fields for Statistical Characterization of Brain Sub-Manifolds. Ph.D. thesis, Washington University, Saint Louis, Missouri.
Khan, A.R., Beg, M.F., 2008. Representation of time-varying shapes in the large deformation diffeomorphic framework. In: 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro 2008 (ISBI 2008).
Krebs, H.A., Henseleit, K., 1932. Untersuchungen über die harnstoffbildung im tierkörper. Hoppe-Seyler's Zeitschrift für Physiol. Chemie. 210, 33–66.
Kuo, H.H., 1975. Gaussian Measures in Banach Spaces. Lecture Notes in Mathematics, vol. 463. Springer-Verlag, Berlin, Heidelberg, New York.
Langner, U.W., Keall, P.J., 2009. Accuracy in the localization of thoracic and abdominal tumors using respiratory displacement, velocity, and phase. Med. Phys. 36, 386–393.
Li, T., Koong, A., Xing, L., 2007. Enhanced 4D cone-beam CT with inter-phase motion model. Med. Phys. 34, 3688–3695.
Lu, W., Mackie, T.R., 2002. Tomographic motion detection and correction directly in sinogram space. Phys. Med. Biol. 47, 1267–1284.
Mair, B.A., Gilland, D.R., Sun, J., 2006. Estimation of images and nonrigid deformations in gated emission ct. IEEE Trans. Med. Imag. 25, 1130–1144.
Miller, M.I., Younes, L., 2001. Group actions, homeomorphisms, and matching: a general framework. Int. J. Comp. Vis. 41, 61–84.
Pan, T., Lee, T.Y., Rietzel, E., Chen, G.T.Y., 2004. 4D-CT imaging of a volume influenced by respiratory motion on multi-slice CT. Med. Phys. 31, 333.
Pevsner, A., Davis, B., Joshi, S., Hertanto, A., Mechalakos, J., Yorke, E., Rosenzweig, K., Nehmeh, S., Erdi, Y.E., Humm, J.L., Larson, S., Ling, C.C., Mageras, G.S., 2006. Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated CT images. Med. Phys. 33, 369–376.
Prince, J.L., Links, J.M., 2006. Medical Imaging Signals and Systems. Prentice-Hall.
Reyes, M., Malandain, G., Koulibaly, P.M., González-Ballester, M.A., Darcourt, J., 2007. Model-based respiratory motion compensation for emission tomography image reconstruction. Phys. Med. Biol. 52, 3579.
Rietzel, E., Pan, T., Chen, G.T.Y., 2005. Four-dimensional computed tomography: image formation and clinical protocol. Med. Phys. 32, 874–889.
Rit, S., Wolthaus, J., van Herk, M., Sonke, J.J., 2008. On-the-fly motion-compensated cone-beam CTusing an a priori motion model. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer-Verlag, pp. 729–736.
Rohlfing, T., Calvin R. Maurer, J., Bluemke, D.A., Jacobs, M.A., 2003. Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint. IEEE Trans. Med. Imag. 22, 730–741.
Ruan, D., Esedoglu, S., Fessler, J.A., 2009. Discriminative sliding preserving regularizationin medical image registration. In: 6th IEEE International Symposium on Biomedical Imaging (ISBI 2009), pp. 430–433.
Saddi, K.A., Chefd'hotel, C., Cheriet, F., 2007. Large deformation registration of contrast-enhanced images with volume-preserving constraint. In: Proceedings of International Society for Optical Engineering (SPIE) Conference on Medical Imaging 2007.
Song, S.M., Leahy, R.M., 1991. Computation of 3-D velocity fields from 3-D cine CT images of a human heart. IEEE Trans. Med. Imag. 10, 295–306.
Stam, J., 2001. A simple fluid solver based on the FFT. J. Graph. Tools 6, 383–396.
Szegedi, M., Rassiah-Szegedi, P., Fullerton, G., Salter, B., 2009. SU-FF-J-128: characterization of liver motion based on implanted markers. Med. Phys. 36, 2506.
Szegedi, M., Rassiah-Szegedi, P., Fullerton, G., Wang, B., Salter, B., 2010. A proto-type design of a real-tissue phantom for the validation of deformation algorithms and 4d dose calculations. Phys. Med. Biol. 55, 3685–3699.
Trouvé, A., 1995. An infinite dimensional group approach for physics based models in patterns recognition.
Vardi, Y., Lee, D., 1993. From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems. J. R. Stat. Soc. B 55, 569–612.
Vedam, S.S., Keall, P.J., Kini, V.R., Mostafavi, H., Shukla, H.P., Mohan, R., 2003. Acquiring a four-dimensional computed tomography dataset using an external respiratory signal. Phys. Med. Biol. 48, 45–62.
Wilson, D.W., Tsui, B.M.W., 1993. Noise properties of filtered-backprojection and ML-EM reconstructed emission tomographic images. IEEE Trans. Nucl. Sci. 40, 1198–1203.
Yu, H., Wang, G., 2007. Data consistency based rigid motion artifact reduction in fan-beam CT. IEEE Trans. Med. Imag. 26, 249–260.
Zeng, R., Fessler, J.A., Balter, J.M., 2007. Estimating 3-D respiratory motion from orbiting views by tomographic image registration. IEEE Trans. Med. Imag. 26, 153–163.

CHAPTER 3

AUTOCALIBRATING CT IMAGE RECONSTRUCTION

The previous chapter contained a method for estimating organ motion during image reconstruction, assuming scanner geometry is accurately known. In this chapter, a related problem is considered, in which organ motion is presumed to be minimal (for instance, the skull may be immobilized), while the scanner geometry is poorly calibrated. As in the case of 4D image reconstruction, geometric parameters of the scan are estimated during image formation. However, in this case, only scanner geometry and a 3D static image are estimated, while the subject's anatomy is assumed to be stationary.

Two-dimensional mobile C-arm fluoroscopic x-ray imaging is a widespread clinical tool due to its low cost of deployment and ability to produce realtime 2D imagery. Because of this, there is a real opportunity to push 3D medical imaging into previously infeasible applications. However, since most C-arm fluoroscopes are not designed for accurate 3D volumetric imaging, forming 3D images from the obtained projection data is difficult.

One of the limitations of using mobile C-arm for 3D conebeam CT image reconstruction is the lack of precise geometric calibration information about the individual 2D projections. In a typical stationary 3D CT scanner, gantry rotation and projection timing are automated, both strictly controlled and calibrated. Many C-arm fluoroscopes do not have motorized gantries, and must be positioned manually by the clinician. In order to acquire sufficient data for 3D image reconstruction, many projections must be acquired by manually swinging the gantry through a significant angular range of between 100 and 200 degrees. Positioning is complicated by possible lack of degrees of freedom in the C-arm gantry, and by *intrinsic* perspective distortion due to gantry flex and sag.

We have developed a method to use the images acquired from conventional C-arm without any external calibration devices to reconstruct a three-dimensional image. This is accomplished by acquiring a series of projections using the C-arm in an imprecisely determined orientation, then jointly estimating the pose calibration and iteratively recon-

structing the image. The automatic calibration technique estimates the mobile C-arm geometry using the raw data and reconstructs the image using an alternating iterative expectation-maximization (EM) technique.

3.1 C-Arm Fluoroscope Scanner Geometry

In order to describe C-arm fluorscope geometry, we first introduce some notation. The position of a point in space is described by its *world coordinates*, $p = (x, y, z) \in \mathbb{R}^3$. World coordinates describe points in space relative to some fixed origin. For instance, the patient is assumed to be stationary so that points in the body are given by unchanging world coordinates. The reconstructed image $I \in L^2(\mathbb{R}^3)$ contains voxels representing x-ray attenuation in world coordinates.

We assume that the x-ray detector is planar, and is fixed some distance $f \in \mathbb{R}$ (called the *source-to-image distance* or *SID*) from the x-ray source as shown in Fig. 3.1. Points within the planar detector are described by the *projection coordinates* $(u, v) \in \mathbb{R}^2$.

In order to relate projection coordinates with world coordinates and formalize the projection operation, we introduce another 3D coordinate frame called *camera coordinates* and denoted by $\boldsymbol{p}' = (x', y', z') \in \mathbb{R}^3$. In camera coordinates, the x-ray source is at the

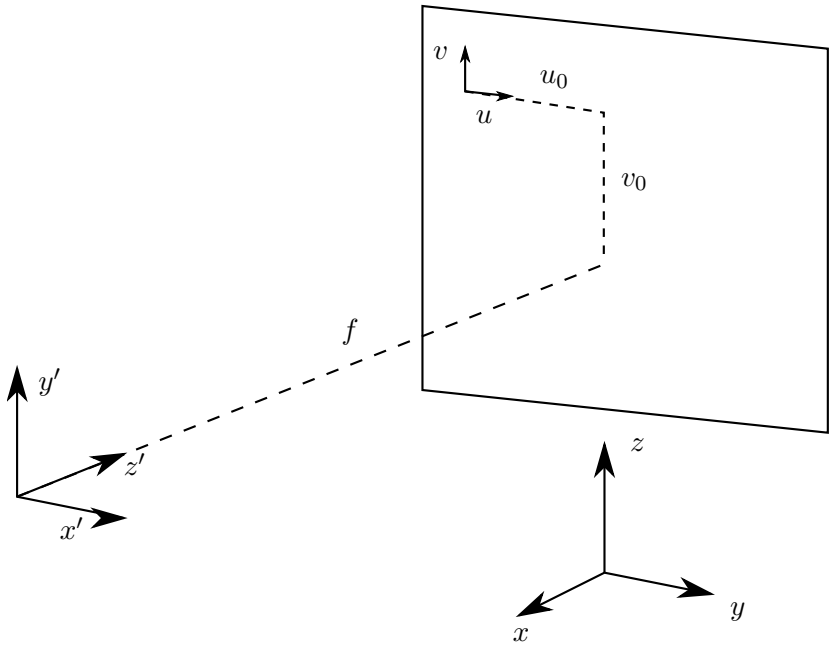


Figure 3.1. Camera geometry for flat-panel detector. Notice that the world origin x, y, z does not correspond to the camera origin x', y', z', and that z' points directly toward the detector.

origin $(0,0,0)$, and the positive z' axis points directly to the detector plane. In camera coordinates, the point $(0,0,f)$ represents the closest point between the x-ray source and the detector plane. We denote this point in projection coordinates by (u_0, v_0), and refer to it as the *piercing point*.

The u direction (horizontal) in projection coordinates corresponds to the x' direction in camera coordinates. Together, the projection and camera coordinates define an imaging system independent of orientation in the real world. That is, given an attenuation function $I'(x', y', z')$, which is given in camera coordinates, the total attenuation $P\{I'\}(u, v)$ of the intensity detected at projection point (u, v) is given by the line integral of I':

$$P\{I\}(u,v) = \sqrt{(u-u_0)^2 + (v-v_0)^2 + f^2} \int_0^1 I'\left(s(u-u_0), s(v-v_0), sf\right) ds. \tag{3.1}$$

Note that the factor outside the integral turns the integral into a proper line integral from $(0,0,0)$ to $(u-u_0, v-v_0, f)$ by multiplying by the line segment length. For clarity, we use the notation γ for this length factor:

$$\gamma(u, v; u_0, v_0, f) := \sqrt{(u-u_0)^2 + (v-v_0)^2 + f^2}. \tag{3.2}$$

As camera coordinates are simply another inertial reference frame, they can be related to world coordinates via some rigid transformation consisting of a rotation and a translation. Various representations of such transformations are in common use, including homogeneous matrices [4] and pairs of translations (vectors in $\mathbb{R}^3$) along with rotations described by either orthogonal matrices, axis-angle representations, or unit quaternions [3]. In this work, I represent a rigid transformation as a pair $(R, \boldsymbol{T}) \in \mathrm{SO}(3) \times \mathbb{R}^3$ and I define its action on a point $\boldsymbol{p}$ in $\mathbb{R}^3$ as

$$(R, \boldsymbol{T}).\boldsymbol{p} = R\boldsymbol{p} + \boldsymbol{T}, \tag{3.3}$$

where the lower dot denotes the group action given by this definition and $R\boldsymbol{p}$ is the usual matrix-vector multiplication. The rotation R is represented as a 3-by-3 orthogonal matrix and, since $R^T R = id$, the identity matrix, the inverse transformation is given by the transformation $(R^T, -R^T\boldsymbol{T})$.

We use $(R, \boldsymbol{T})$ to denote the transformation used to convert an attenuation image to camera coordinates, so that world coordinates are related to camera coordinates by

$$\boldsymbol{p} = (R, \boldsymbol{T}).\boldsymbol{p}'. \tag{3.4}$$

Thus given an image $I(\boldsymbol{p})$, defined in world coordinates, its corresponding camera coordinate image is

$$I'(\boldsymbol{p}') = I(\boldsymbol{p}) = I((R, \boldsymbol{T}).\boldsymbol{p}') = I(R\boldsymbol{p}' + \boldsymbol{T}). \tag{3.5}$$

Given the SID f, piercing point (u_0, v_0), and transformation $(R, \boldsymbol{T})$, which we refer to as *pose*, the projection P of a world-coordinate image I is

$$P\{I\}(u,v) = \gamma(u,v;u_0,v_0,f) \int_0^1 I'\left(s(u-u_0), s(v-v_0), sf\right) ds \tag{3.6}$$

$$= \gamma(u,v;u_0,v_0,f) \int_0^1 I\left(R(s(u-u_0), s(v-v_0), sf) + \boldsymbol{T}\right) ds. \tag{3.7}$$

As mentioned, the parameters that determine the projection completely are the pose $(R, \boldsymbol{T})$, the SID f, and the piercing point (u_0, v_0). Typically, the pose parameters, rotation and translation, are referred to as *extrinsic parameters*, while the piercing point and SID are called *intrinsic parameters*. In the case of C-arm fluoroscopy, intrinsic parameters are more precisely specified than are extrinsic parameters. The reason for this is that typically, the pose is determined by manual rotation of the gantry, which may rotate or bounce unintentionally due to the mechanical design of the scanner. This unknown perturbation usually effects the entire gantry, so that the intrinsic parameters, representing the relation of some parts of the gantry to others, is effected to a much smaller degree than the extrinsic parameters relating the gantry to the subject, to which it is not mechanically coupled.

3.2 Expectation-Maximization Image Reconstruction

In this section, I derive an expectation-maximization method reconstruction of an image I using a Poisson noise model. The reconstructed image I represents the x-ray attenuation coefficient at every point in space. Given an x-ray source at a point $x_w \in \mathbb{R}^3$ emitting x-rays with intensity I_0 and a detector element such as a CCD pixel placed behind a phosphor screen at a point. Detector noise is Poisson distributed around the x-ray intensity $I_0 \exp(P_j\{I\})$. Image reconstruction under a Poisson noise model is typically optimized using expectation maximization (EM) [2].

Expectation-maximization image reconstruction has been very well studied. The most common algorithm for CT EM reconstruction involves the so-called *Richardson-Lucy iterations* (c.f. White [5]):

$$I(x) \mapsto \frac{I(x)}{\sum_j P_j^\dagger\{\mathbf{1}\}(x)} \sum_j P_j^\dagger \left(\frac{d_j}{P_j\{I\}} \right). \tag{3.8}$$

In the above formula, the data are 2D projection images d_j, $P_j^\dagger$ is the backprojection operator defined as the adjoint of the conebeam projection, and $\mathbf{1}$ is a 2D projection image consisting of all ones. These multiplicative image updates are repeatedly computing, and are guaranteed to increase the likelihood function. Thus, while there is no explicit step-size

parameter in EM CT reconstruction as there is with Gaussian gradient descent, image quality is often determined by the number of computed iterations.

3.2.1 Total Variation

I have implemented the total variation (TV) regularization scheme presented by Panin, Zeng, and Gullberg [1] for use in MAP EM reconstruction. The objective function used in this case is a sum of the Poisson data log-likelihood, with a regularization term

$$U(I) = \int |\nabla I(x)| dx. \tag{3.9}$$

The essential element of Panin and Gullberg's method is what is referred to as the one-step-late (OSL) total variation update. As derived in [1, Eq. 8], this update amounts to an additional multiplicative factor within the Richardson-Lucy image updates. For the OSL-TV algorithm, the image update takes the form

$$I(x) \mapsto \frac{I(x)}{\left(\sum_j P^\dagger\{1\}\right)(x) + \lambda \frac{\partial}{\partial I(x)} U(I)} \sum_j P_j^\dagger \left(\frac{d_j}{P_j\{I\}}\right), \tag{3.10}$$

where λ is a scalar parameter controlling the strength of the TV regularization term. A significant advantage of the OSL total variation method is that it retains the desireable lack of stepsize parameter, similar to unregularized ML-EM. As is shown in detail by Panin and Gullberg, the update above can be computed in a numerically accurate fashion using an approprate stencil for the term $\frac{\partial}{\partial I(x)} U(I)$.

3.3 Optimization of Projection Parameters

The EM conebeam reconstruction algorithm, assuming perfect calibration, iteratively updates a 3D image estimate by comparing its projections against the observed data and backprojecting the discrepancy. The autocalibrating algorithm, at each iteration of the EM algorithm, also updates the estimated calibration parameters. We use a fixed step size gradient descent optimization scheme, but this could be replaced by any gradient-based algorithm. In this chapter, gradients with respect to geometric parameters, both intrinsic and extrinsic, are derived.

3.3.1 Parameter Gradient Computation

The steepest descent scheme depends on computation of the derivatives of (3.7) with respect to all projection parameters. Each of these is straightforward, and amounts to an

application of the chain rule. First consider the simplest parameter gradient: that with respect to translation $\boldsymbol{T}$:

$$\frac{\partial}{\partial \boldsymbol{T}} P\{I\}(u,v) = \gamma(u,v;u_0,v_0,f) \int_0^1 (\nabla I)\left(R(s(u-u_0), s(v-v_0), sf) + \boldsymbol{T}\right) ds. \tag{3.11}$$

Notice that this is a line integral of the *gradient* of I. This line integral could be computed by precomputing the gradient of I then integrating each component. However, as the image is defined on a discrete grid and the line integral approximated via a ray-marching and interpolation scheme, I implement this line integral using a function that interpolates not only the image value but the trilinear interpolation function. This has the advantage of using nearly the same amount of memory as the ordinary projection computation, which is important for implementation on limited-memory GPU hardware.

The gradient with respect to R is slightly more complicated because R is constrained to be a rotation matrix. We model a perturbation of R as a small rotation applied on the left to R. That small rotation is parametrized by a skew-symmetric matrix $S = W - W^T$, which is exponentiated to generate a true rotation matrix:

$$R \mapsto \exp(W - W^T)R. \tag{3.12}$$

The matrix exponential of a skew-symmetric matrix is necessarily a rotation matrix, so this guarantees that whatever matrix W is used, the resulting perturbation of R remains a rotation. Now we use the Fréchet variation to compute the gradient of $P\{I\}$ with respect to R:

$$\left\langle \frac{\partial}{\partial R} P\{I\}(u,v), W \right\rangle_{Frob} = \frac{d}{d\epsilon}\Big|_{\epsilon=0} P\{I; \exp(\epsilon(W - W^T))R\}(u,v). \tag{3.13}$$

Again, the right-hand side is computed using the chain rule, along with the fact that

$$\frac{d}{d\epsilon} \exp(\epsilon(W - W^T))R = (W - W^T)\exp(\epsilon(W - W^T))R. \tag{3.14}$$

Equation (3.13) then becomes

$$\left\langle \frac{\partial}{\partial R} P\{I\}(u,v), W \right\rangle_{Frob} = \gamma(u,v;u_0,v_0,f) \int_0^1 (\nabla I)\left(R\boldsymbol{p}'(s) + \boldsymbol{T}\right)^T (W - W^T) R\boldsymbol{p}'(s) ds, \tag{3.15}$$

where I have used the notation $\boldsymbol{p}'(s) = (s(u-u_0), s(v-v_0), sf)$. I use the following identity relating the right-hand side above to the Frobenius inner product:

$$\boldsymbol{w}^T A \boldsymbol{v} = \langle \boldsymbol{w}\boldsymbol{v}^T, A \rangle_{Frob} \tag{3.16}$$

to write

$$\left\langle \frac{\partial}{\partial R} P\{I\}(u,v), W \right\rangle = \gamma(u,v;u_0,v_0,f) \int_0^1 \left\langle (\nabla I)\left(R\boldsymbol{p}'(s)+\boldsymbol{T}\right)\left(R\boldsymbol{p}'(s)\right)^T, W - W^T \right\rangle ds. \tag{3.17}$$

Rearranging this, we find that the gradient with respect to R is given by the skew-symmetric matrix

$$\frac{\partial}{\partial R} P\{I\}(u,v) = Y - Y^T, \tag{3.18}$$

where Y is computed using the formula

$$Y := \gamma(u,v;u_0,v_0,f) \int_0^1 (\nabla I)\left(R\boldsymbol{p}'(s)+\boldsymbol{T}\right)\left(R\boldsymbol{p}'(s)\right)^T ds. \tag{3.19}$$

This is similar to the gradient with respect to $\boldsymbol{T}$, except that instead of integrating the gradient, the outer product of the gradient and the rotated point $R.\boldsymbol{p}'(s)$ is integrated. Some simplification can be accomplished in order to increase the efficiency of this computation. First notice that the skew-symmetrization of an outer product, $\boldsymbol{w}\boldsymbol{v}^T$, is in fact the skew-symmetric matrix generated by the cross product:

$$\boldsymbol{w}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{w}^T = *(\boldsymbol{w}\times\boldsymbol{v}), \tag{3.20}$$

where the star indicates the standard mapping from $\mathbb{R}^3$ to $\mathfrak{so}(3)$, as discussed in more detail in Sec. 4.6.2 (page 49). Using this fact, the gradient with respect to R may be written as a 3-vector given by integration of these cross products:

$$\frac{\partial}{\partial R} P\{I\}(u,v) = \gamma(u,v;u_0,v_0,f) \int_0^1 (\nabla I)\left(R\boldsymbol{p}'(s)+\boldsymbol{T}\right) \times \left(R\boldsymbol{p}'(s)\right) ds. \tag{3.21}$$

In order to perform a gradient descent step in a direction given by a 3-vector $\boldsymbol{v}$, Rodrigues' formula is applied:

$$R \mapsto R + \sin\|\boldsymbol{v}\|(*\boldsymbol{v})R + (1-\cos\|\boldsymbol{v}\|)(\boldsymbol{v}\boldsymbol{v}^T R - R), \tag{3.22}$$

which corresponds to the exponential map in SO(3) in the direction $*\boldsymbol{v} \in \mathfrak{so}(3)$.

3.3.1.1 Intrinsic Parameter Derivatives

Derivatives with respect to the intrinsic parameters u_0, v_0, and f are computed in a similar fashion. For simplicity, I combine them into a single vector $\boldsymbol{\iota} = (u_0, v_0, -f)$ representing the projection coordinate origin in camera coordinates, so that

$$\boldsymbol{p}'(s) = s((u,v,0) - \boldsymbol{\iota}) \tag{3.23}$$

$$\frac{\partial}{\partial \boldsymbol{\iota}} \boldsymbol{p}'(s) = \mathrm{diag}(s). \tag{3.24}$$

This simply means the Jacobian matrix of $\boldsymbol{p}'(s)$ with respect to derivatives in the intrinsic parameters ι is s times the identity matrix.

The gradient of $P\{I\}$ with respect to ι is computed using the product rule. The first term is the gradient of γ multiplied by the line integral and the second term is γ times the gradient of the line integral. Notice that in this notation, γ is given by

$$\gamma(u, v; \boldsymbol{\iota}) = \|\boldsymbol{p}'(1)\| = \|(u, v, 0) - \boldsymbol{\iota}\|, \tag{3.25}$$

so the gradient in the $\boldsymbol{\iota}$ direction is simply the unit vector pointing from $\boldsymbol{\iota}$ to the point $(u, v, 0)$:

$$\frac{\partial}{\partial \boldsymbol{\iota}}\gamma(u, v; \boldsymbol{\iota}) = \frac{(u, v, 0) - \boldsymbol{\iota}}{\|(u, v, 0) - \boldsymbol{\iota}\|}. \tag{3.26}$$

Then we can write the gradient with respect to intrinsic parameters:

$$\frac{\partial}{\partial \boldsymbol{\iota}}P\{I\}(u, v) = \frac{(u, v, 0) - \boldsymbol{\iota}}{\|(u, v, 0) - \boldsymbol{\iota}\|}P\{I\}(u, v) + \|(u, v, 0) - \boldsymbol{\iota}\| \int_0^1 sR^T(\nabla I)\left(R\boldsymbol{p}'(s) + \boldsymbol{T}\right) ds. \tag{3.27}$$

Notice the factor of s in the integrand, and the multiplication by R^T which comes from the chain rule in this gradient with respect to ι. That matrix multiplication can be pulled out of the integral. Also note that

$$s\|(u, v, 0) - \boldsymbol{\iota}\| = \|\boldsymbol{p}'(s)\|, \tag{3.28}$$

which further simplifies the above expression to

$$\frac{\partial}{\partial \boldsymbol{\iota}}P\{I\}(u, v) = \frac{(u, v, 0) - \boldsymbol{\iota}}{\|(u, v, 0) - \boldsymbol{\iota}\|}P\{I\}(u, v) + R^T \int_0^1 \|\boldsymbol{p}'(s)\|(\nabla I)\left(R\boldsymbol{p}'(s) + \boldsymbol{T}\right) ds. \tag{3.29}$$

In order to compute the derivatives with respect to all parameters, the following quantities must be integrated over s from zero to one:

$$I(R.\boldsymbol{p}'(s) + \boldsymbol{T}) \tag{3.30}$$

$$(\nabla I)(R\boldsymbol{p}'(s) + \boldsymbol{T}) \tag{3.31}$$

$$(\nabla I)(R\boldsymbol{p}'(s) + \boldsymbol{T}) \times (R.\boldsymbol{p}'(s)) \tag{3.32}$$

$$\|\boldsymbol{p}'(s)\|(\nabla I)(R\boldsymbol{p}'(s) + \boldsymbol{T}). \tag{3.33}$$

This scalar and three vector quantities should be integrated simultaneously so as to avoid repeated interpolation.

3.4 Results

We analyze the performance of this algorithm on both synthetic and real C-arm projection data.

3.4.1 Skull Analytic Phantom

We performed a simulated phantom study using a public domain head CT dataset from the "University of North Carolina Volume Rendering Test Data Set" (available from `http://www-graphics.stanford.edu/data/voldata/`), shown in Fig. 3.2. We generated a perturbed angular sampling, simulating projections at angles which were randomly perturbed from a nominal trajectory of one projection per degree. We reconstructed images using these perturbed angles, simulating a situation in which the poses were irregular but measured. This "best case" image reconstruction exhibited faint streaking artifacts due to irregular angular sampling, as seen in Fig. 3.2. In order to overcome the artifacts in our best-case reconstruction, we added a one-step-late total variation (OSL-TV) regularization step to our EM algorithm. The updated algorithm ran at nearly the same speed as the original EM algorithm, but was shown to eliminate streaking. These image reconstructions were found to converge within 5–10 iterations.

Next, I tested whether we could simultaneously estimate pose while reconstructing the image. I implemented the fixed step-size gradient descent scheme, interleaved with EM steps, which alternately optimizes the pose estimate and updates the image. We tested this 3D code using 223 projections (180 degrees plus cone angle) at poses which were perturbed from a regular 1 degree separation by random rotations within ± 5 degrees and ± 5 mm in translation in each direction. The algorithm was shown to produce a reconstruction nearly indistinguishable from the "best-case" reconstruction, as shown in Fig. 3.3. As the algorithm used a gradient descent to estimate pose in addition to the EM image reconstruction steps, the algorithm is slower than ordinary image reconstruction. Whereas the OS-EM+TV image reconstruction algorithm converged within 10 iterations, the pose+image reconstruction

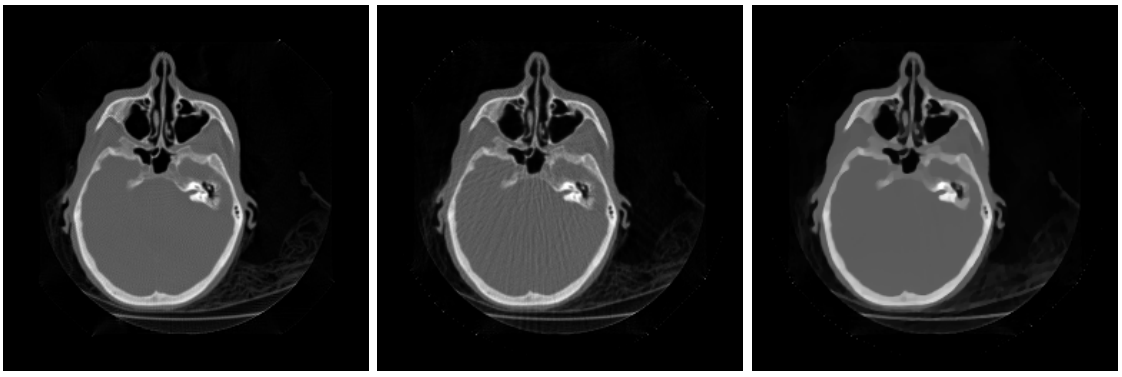


Figure 3.2. Head CT dataset used as phantom (left), EM best-case reconstruction without total variation (center), and EM best-case reconstruction with TV (right).

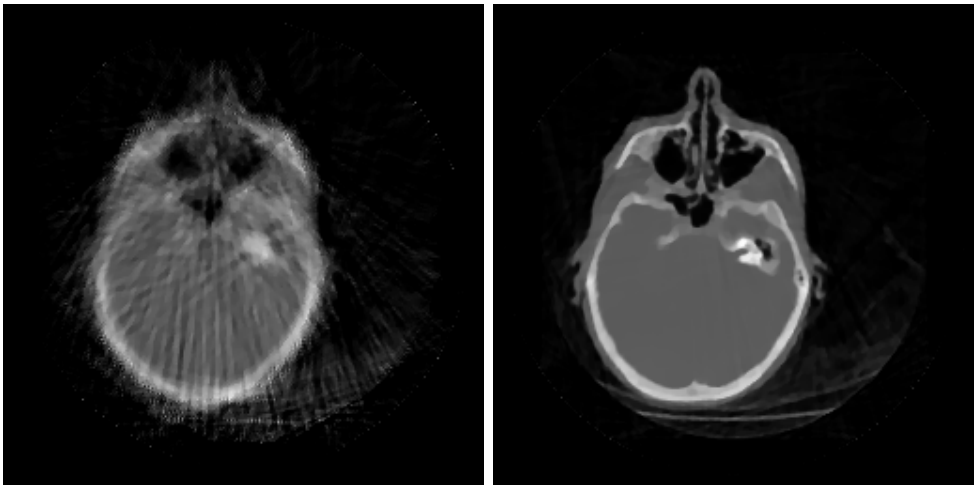


Figure 3.3. Uncalibrated reconstruction with TV (left) and auto-calibrated reconstruction with TV (right). Notice that the autocalibrated reconstruction looks nearly identical to the best-case reconstruction.

algorithm took roughly 30 iterations to converge.

We suspected that the results we obtained using full angular range may be significantly poorer when using a limited angular range. In order to test the effect of the angular range, we ran a test using the 3D code in which only 140 projections were obtained, perturbed as before from a regular 1 degree spacing. The resulting best-case reconstruction showed, as expected, some smearing artifacts in the undetermined angles, but total variation was shown to improve this image, as seen in Fig. 3.4. More importantly, the limited angular range was not shown to degrade the performance of pose estimation, and we observed that our auto-calibrated reconstruction result was again nearly indistinguishable from the best case reconstruction.

3.4.2 Skull Turntable Phantom

The physical phantom data consisted of a skull phantom, mounted on a turntable and imaged using a fixed C-arm fluoroscope. The algorithm used was the same as that in the previous section, but the data were acquired from a real detector, instead of being numerically simulated.

Notice that in this case, although the turntable provides a nominal angular spacing of one degree, manual alignment of the turntable and fluoroscope make it impossible to have a ground-truth trajectory. Instead, in Fig. 3.5, we show the nominal reconstruction, using the ideally-aligned trajectory parameters, which exhibits substantial ghosting artifacts.

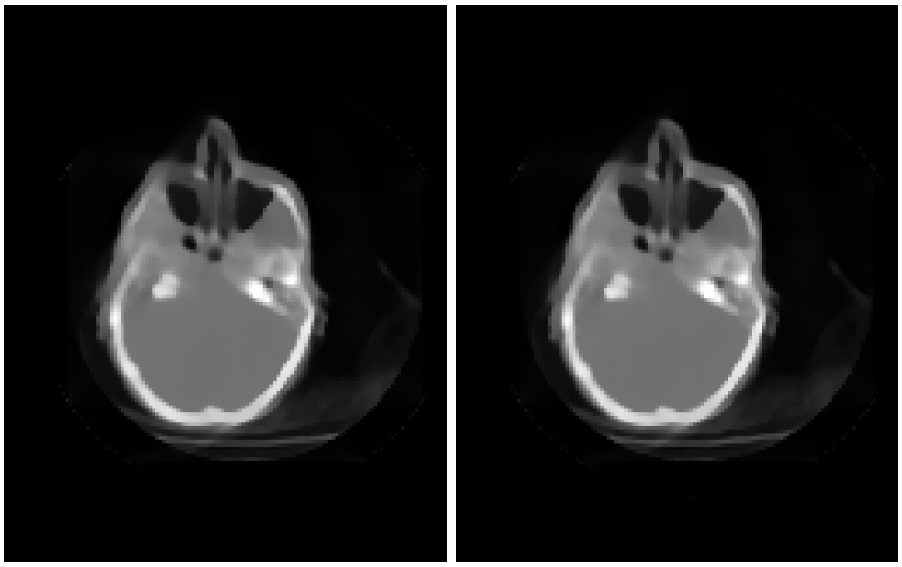


Figure 3.4. Best-case reconstruction with TV and 140 degree angular range, and autocalibrated reconstruction with TV on same data. Note that this was run on a lower resolution due to longer runtimes involved in 3D studies.

The autocalibrating algorithm is able to correct these artifacts and provide an accurate reconstruction of the skull phantom.

3.5 Conclusion

We have developed a technique which takes 2D C-arm fluoroscope projection data along with inaccurate calibration information, and reconstructs a 3D image while estimating optimal calibration parameters. This enables 3D conebeam CT reconstruction in applications such as C-arm fluoroscopy, which without autocalibration yield reconstructed images so heavily corrupted by artifacts that they are unusable. Our results demonstrate that this technique results in a dramatic reduction in image reconstruction artifacts.

Acknowledgments

The author would like to acknowledge the support from GE Healthcare, which made this explorational research possible.

References

[1] VY Panin, GL Zeng, and GT Gullberg. "Total Variation Regulated EM Algorithm [SPECT Reconstruction]". In: *Nuclear Science, IEEE Transactions on* 46.6 (1999), pp. 2202–2210.

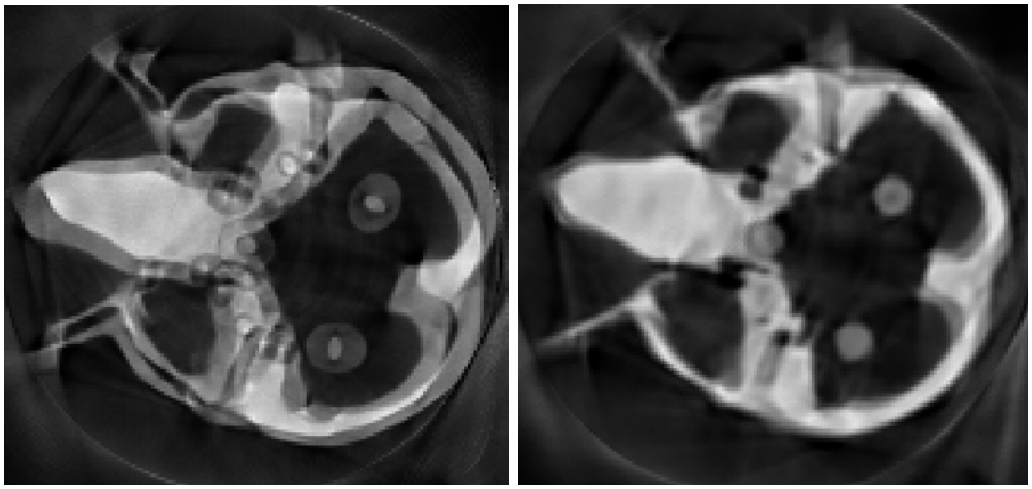


Figure 3.5. Turntable skull phantom reconstructed using OS-EM with no total variation and the nominal trajectory (left), and autocalibrated reconstruction (right). Notice the reduction in ghosting artifacts due to correction of misalignment.

[2] Lawrence A Shepp and Yehuda Vardi. "Maximum Likelihood Reconstruction for Emission Tomography". In: *Medical Imaging, IEEE Transactions on* 1.2 (1982), pp. 113–122.

[3] Ken Shoemake. "Animating Rotation with Quaternion Curves". In: *ACM SIGGRAPH Computer Graphics* 19.3 (1985), pp. 245–254.

[4] Daniel W VanArsdale. "Homogeneous Transformation Matrices for Computer Graphics". In: *Computers & graphics* 18.2 (1994), pp. 177–191.

[5] Richard L White. "Image Restoration Using the Damped Richardson-Lucy Method". In: *The Restoration of HST Images and Spectra II* (1994), pp. 104–110.

CHAPTER 4

POLYNOMIAL REGRESSION ON RIEMANNIAN MANIFOLDS

The following chapter consists of a manuscript, published (Open Access) in the Journal of Mathematical Imaging and Vision (Springer), that details a method of fitting Riemannian polynomial curves to data lying in Riemannian manifolds. Gradient-based optimization using the adjoint method for gradient computation is derived in the settings of general Riemannian manifolds, Lie groups with right (and left) invariant Riemannian metrics, and in homogeneous spaces acted upon by such Lie groups. The adjoint differential equations that determine gradient computation generalize the famous Jacobi equation, just as Riemannian polynomials generalize geodesics to higher order. Riemannian polynomials are shown to provide improved curve fitting in a number of studies applied to real data. The reader is encouraged to consult Appendix A for a more through treatment of geodesic regression in Lie groups, and Appendix B for a similar treatment of Lie group actions, momentum maps, and Lie squares regression in these cases.

Reprinted from Journal of Mathematical Imaging and Vision September 2014, Volume 50, Issue 1, pp 32–52 with permission from Springer

J Math Imaging Vis (2014) 50:32–52
DOI 10.1007/s10851-013-0489-5

Intrinsic Polynomials for Regression on Riemannian Manifolds

Jacob Hinkle · P. Thomas Fletcher · Sarang Joshi

Received: 11 February 2013 / Accepted: 28 December 2013 / Published online: 22 February 2014

Abstract We develop a framework for polynomial regression on Riemannian manifolds. Unlike recently developed spline models on Riemannian manifolds, Riemannian polynomials offer the ability to model parametric polynomials of all integer orders, odd and even. An intrinsic adjoint method is employed to compute variations of the matching functional, and polynomial regression is accomplished using a gradient-based optimization scheme. We apply our polynomial regression framework in the context of shape analysis in Kendall shape space as well as in diffeomorphic landmark space. Our algorithm is shown to be particularly convenient in Riemannian manifolds with additional symmetry, such as Lie groups and homogeneous spaces with right or left invariant metrics. As a particularly important example, we also apply polynomial regression to time-series imaging data using a right invariant Sobolev metric on the diffeomorphism group. The results show that Riemannian polynomials provide a practical model for parametric curve regression, while offering increased flexibility over geodesics.

Keywords Polynomial · Riemannian geometry · Regression · Rolling maps · Lie groups · Shape space

1 Introduction

Comparative studies are essential to biomedical statistical analysis. In the context of shape, such analyses are used to discriminate between healthy and disease states based on observations of anatomical shapes within individuals in the two populations [35]. Commonly, in these methods the shape data are modelled on a Riemannian manifold and intrinsic coordinate-free manifold-based methods are used [8]. This prevents bias due to arbitrary choice of coordinates and avoids the influence of unwanted effects. For instance, by modelling shapes with a representation incapable of representing scale and rotation of an object and using intrinsic manifold-based methods, scale and rotation are guaranteed not to effect the analysis [19].

Many conditions such as developmental disorders and neurodegeneration are characterized not only by shape characteristics, but by abnormal *trends* in anatomical shapes over time. Thus it is often the temporal dependence of shape that is most useful for comparative shape analysis. The field of regression analysis involves studying the connection between independent variables and observed responses [34]. In particular, this includes the study of temporal trends in a observed data.

In this work, we extend the recently developed geodesic regression model [12] to higher order polynomials using intrinsic Riemannian manifold-based methods. We show that this Riemannian polynomial model is able to provide increased flexibility over geodesics, while remaining in the parametric regression setting. The increase in flexibility is particularly important, as it enables a more accurate description of shape trends and, ultimately, more useful comparative regression analysis.

While our primary motivation is shape analysis, the Riemannian polynomial model is applicable in a variety of applications. For instance, directional data is commonly modelled as points on the sphere $\mathbb{S}^2$, and video sequences representing human activity are modelled in Grassmannian manifolds [36].

In computational anatomy applications, the primary objects of interest are elements of a group of symmetries acting

J. Hinkle ( ) · P.T. Fletcher · S. Joshi
SCI Institute, University of Utah, 72 Central Campus Dr., Salt Lake City, Utah 84112, USA
e-mail: jacob@sci.utah.edu

on the space of observable data. For instance, rigid motion is studied using the groups SO(3) and SE(3), acting on a space of landmark points or scalar images. Non-rigid motion and growth is modelled using infinite-dimensional diffeomorphism groups, such as in the currents framework [37] for unlabelled landmarks or the large deformation diffeomorphic metric mapping (LDDMM) framework of deforming images [30]. We show that in the presence of a group action, optimization of our polynomial regression model using an adjoint method is particularly convenient.

This work is an extension of the Riemannian polynomial regression framework first presented by Hinkle et al. [15]. In Sects. 5–7, we give a new derivation of polynomial regression for Lie groups and Lie group actions with Riemannian metrics. By performing the adjoint optimization directly in the Lie algebra, the computations in these spaces are greatly simplified over the general formulation. We show how this Lie group formulation can be used to perform polynomial regression on the space of images acted on by groups of diffeomorphisms.

1.1 Regression Analysis and Curve-Fitting

The study of the relationship between measured data and descriptive variables is known as the field of regression analysis. As with most statistical techniques, regression analyses can be broadly divided into two classes: parametric and non-parametric. The most widely used parametric regression methods are linear and polynomial regression in Euclidean space, wherein a linear or polynomial function is fit in a least-squares fashion to observed data. Such methods are the staple of modern data analysis. The most common non-parametric regression approaches are kernel-based methods and spline smoothing approaches which provide great flexibility in the class of regression functions. However, their non-parametric nature presents a challenge to inference problems; if, for example, one wishes to perform a hypothesis test to determine whether the trend for one group of data is significantly different from that of another group.

In previous work, non-parametric kernel-based and spline-based methods have been extended to observations that lie on a Riemannian manifold with some success [8, 18, 22, 26], but intrinsic parametric regression on Riemannian manifolds has received limited attention. Recently, Fletcher [12] and Niethammer et al. [31] have each independently developed a form of parametric regression, geodesic regression, which generalizes the notion of linear regression to Riemannian manifolds. Geodesic models are useful, but are limited by their lack of flexibility when modelling complex trends.

Fletcher [12] defines a geodesic regression model by introducing a manifold-valued random variable Y,

$$Y = \operatorname{Exp}\big(\operatorname{Exp}(p, Xv), \epsilon\big), \tag{1}$$

where $p \in M$ is an initial point and $v \in T_pM$ an initial velocity. The geodesic curve $\operatorname{Exp}(p, Xv)$ then relates the independent variable $X \in R$ to the dependent random variable Y, via this equation and the Gaussian random vector $\epsilon \in T_{\operatorname{Exp}(p,Xv)}M$. In this paper, we extend this model to a polynomial regression model

$$Y = \operatorname{Exp}\big(\gamma(X), \epsilon\big), \tag{2}$$

where the curve $\gamma(X)$ is a Riemannian polynomial of integer order k. In the case that M is Euclidean space, this model is simply

$$Y = p + \sum_{i=1}^{k} \frac{v_i}{i!} X^i + \epsilon, \tag{3}$$

where the point p and vectors v_i constitute the parameters of our model.

In this work we use the common term *regression* to describe methods of fitting polynomial curves using a sum of squared error penalty function. In Euclidean spaces, this is equivalent to solving a maximum likelihood estimation problem using a Gaussian noise model for the observed data. In Riemannian manifolds, the situation is more nuanced, as there is no consensus on how to define Gaussian distributions on general Riemannian manifolds, and in general the least-squares penalty may not correspond to a log likelihood. Many of the examples we will present are symmetric spaces: Kendall shape space in two dimensions, the rotation group, and the sphere, for instance. As Fletcher [12, Sect. 4] explains, least-squares regression in symmetric spaces does, in fact, correspond to maximum likelihood estimation of model parameters, using a natural definition of Gaussian distribution.

1.2 Previous Work: Cubic Splines

Noakes et al. [32] first introduced the notion of Riemannian cubic splines. They fix the endpoints $y_0, y_1 \in M$ of a curve, as well as the derivative of the curve at those points $y_0' \in T_{y_0}M$, $y_1' \in T_{y_1}M$. A Riemannian cubic spline is then defined as any differentiable curve $\gamma : [0, 1] \to M$ taking on those endpoints and derivatives and minimizing

$$\Phi(\gamma) = \int_0^1 \left\langle \nabla_{\frac{d}{dt}\gamma} \frac{d}{dt}\gamma(t), \nabla_{\frac{d}{dt}\gamma} \frac{d}{dt}\gamma(t) \right\rangle dt. \tag{4}$$

As is shown by Noakes et al. [14, 32], between endpoints, cubic splines satisfy the following Euler-Lagrange equation:

$$\nabla_{\frac{d}{dt}\gamma} \frac{d}{dt}\gamma + R\left(\nabla_{\frac{d}{dt}\gamma} \frac{d}{dt}\gamma, \frac{d}{dt}\gamma\right) \frac{d}{dt}\gamma = 0. \tag{5}$$

Cubic splines are useful for interpolation problems on Riemannian manifolds. However, cubic splines provide an

insufficient model for parametric curve regression. For instance, by increasing the order of derivatives in Eq. (4), cubic splines are generalizable to higher order curves. Still, only odd order splines may be defined in this way, and there is no clear way to define even order splines.

Riemannian splines are parametrized by the endpoint conditions, meaning that the space of curves is naturally explored by varying control points. This is convenient if control points such as observed data are given at the outset. However, for parametric curve regression, curve models are preferred that don't depend on the data, such as the initial conditions of a geodesic [12]. Although Eq. (5) provides an ODE which could be used as such a parametric model in a "spline shooting" algorithm, estimating initial position and derivatives as parameters, the curvature term complicates integration and optimization.

1.3 Contributions in This Work

The goal of the current work is to extend the geodesic regression model in order to accommodate more flexibility while remaining in the parametric setting. The increased flexibility introduced by the methods in this manuscript allow a better description of the variability in the data. The work presented in this paper allows one to fit polynomial regression curves on a general Riemannian manifold, using intrinsic methods and avoiding the need for unwrapping and rolling. Since our model includes time-reparametrized geodesics as a special case, information about time dependence is also obtained from the regression without explicit modeling by examining the collinearity of the estimated parameters.

We derive practical algorithms for fitting polynomial curves to observations in Riemannian manifolds. The class of polynomial curves we use, described by Leite & Krakowski [24], is more suited to parametric curve regression than are spline models. These polynomials curves are defined for any integer order and are naturally parametrized via initial conditions instead of control points. We derive explicit formulas for computing derivatives with respect to the initial conditions of these polynomials in a least-squares curve-fitting setting.

In the following sections, we describe our method of fitting polynomial curves to data lying in various spaces. We develop the theory for general Riemannian manifolds, Lie groups with right invariant metrics, and finally for spaces acted on by such Lie groups. In order to keep each application somewhat self-contained, results will be shown in each case in the section in which the associated space is treated, instead of in a separate results section following all the methods.

2 Riemannian Geometry Preliminaries

Before defining Riemannian polynomials, we first review a few basic results from Riemannian geometry and establish a common notation. For a more in-depth treatment of this background material see, for instance, do Carmo [9]. Let (M, g) be a Riemannian manifold. At each point $p \in M$, the metric g defines an inner product on the tangent space T_pM. The metric also provides a method to differentiate vector fields with respect to one another, referred to as the covariant derivative. For smooth vector fields $v, w \in \mathfrak{X}(M)$ and a smooth curve $\gamma : [0, 1] \to M$ the covariant derivative satisfies the following product rule:

$$\frac{d}{dt}\langle v(\gamma(t)), w(\gamma(t))\rangle = \langle \nabla_{\frac{d}{dt}\gamma} v(\gamma(t)), w(\gamma(t))\rangle + \langle v(\gamma(t)), \nabla_{\frac{d}{dt}\gamma} w(\gamma(t))\rangle. \tag{6}$$

A geodesic $\gamma : [0, 1] \to M$ is characterized (for instance) by the conservation of kinetic energy along the curve:

$$\frac{d}{dt}\left\langle \frac{d}{dt}\gamma, \frac{d}{dt}\gamma \right\rangle = 0 = 2\left\langle \nabla_{\frac{d}{dt}\gamma}\frac{d}{dt}\gamma, \frac{d}{dt}\gamma \right\rangle. \tag{7}$$

which leads to the differential equation

$$\nabla_{\frac{d}{dt}\gamma}\frac{d}{dt}\gamma = 0. \tag{8}$$

This is called the geodesic equation and uniquely determines geodesics, parametrized by the initial conditions $(\gamma(0), \frac{d}{dt}\gamma(0)) \in TM$. The mapping from the tangent space at p into the manifold M, defined by integration of the geodesic equation, is called the exponential map and is written $\mathrm{Exp}_p : T_pM \to M$. The exponential map is injective on a zero-centered ball B in T_pM of some non-zero radius. Thus, for a point q within a neighborhood of p, there exists a unique vector $v \in T_pM$ corresponding to a minimal length path under the exponential map from p to q. The mapping of such points q to their associated tangent vectors v at p is called the log map of q at p, denoted $v = \mathrm{Log}_p q$.

Given a curve $\gamma : [0, 1] \to M$, the covariant derivative $\nabla_{\frac{d}{dt}\gamma}$ provides a way to relate tangent vectors at different points along γ. A vector field w is said to be parallel transported along γ if it satisfies the parallel transport equation,

$$\nabla_{\frac{d}{dt}\gamma} w(\gamma(t)) = 0. \tag{9}$$

Notice that the geodesic equation is a special case of parallel transport, under which the velocity is parallel along the curve itself.

3 Riemannian Polynomials

We now introduce Riemannian polynomials as a generalization of geodesics [15]. Geodesics are generalizations to the

Riemannian manifold setting of curves in $\mathbb{R}^d$ with constant first derivative. In the previous section we briefly reviewed how the covariant derivative provides a way to define vector fields which are analogous to constant vector fields along γ, via parallel transport.

We refer to the vector field $\nabla_{\frac{d}{dt}\gamma}\frac{d}{dt}\gamma(t)$ as the acceleration of the curve γ. Curves with parallel acceleration are generalizations of curves in $\mathbb{R}$ whose coordinates are second order polynomials, and satisfy the second order polynomial equation,

$$(\nabla_{\frac{d}{dt}\gamma})^2\frac{d}{dt}\gamma(t)=0. \tag{10}$$

Extending this idea, a cubic polynomial is a curve with parallel jerk (time derivative of acceleration), and so on. Generally, a kth order polynomial in M is defined as a curve $\gamma:[0,1]\to M$ satisfying

$$(\nabla_{\frac{d}{dt}\gamma})^k\frac{d}{dt}\gamma(t)=0 \tag{11}$$

for all times $t\in[0,1]$. As with polynomials in Euclidean space, polynomials are fully determined by initial conditions at $t=0$:

$$\gamma(0)\in M, \tag{12}$$

$$\frac{d}{dt}\gamma(0)\in T_{\gamma(0)}M, \tag{13}$$

$$(\nabla_{\frac{d}{dt}\gamma})^i\frac{d}{dt}\gamma(0)\in T_{\gamma(0)}M,\quad i=1,\ldots,k-1. \tag{14}$$

Introducing vector fields $v_1(t),\ldots,v_k(t)\in T_{\gamma(t)}M$, we write the following system of covariant differential equations, which is equivalent to Eq. (11):

$$\frac{d}{dt}\gamma(t)=v_1(t) \tag{15}$$

$$\nabla_{\frac{d}{dt}\gamma}v_i(t)=v_{i+1}(t),\quad i=1,\ldots,k-1 \tag{16}$$

$$\nabla_{\frac{d}{dt}\gamma}v_k(t)=0. \tag{17}$$

In this notation, the initial conditions that determine the polynomial are $\gamma(0), v_i(0), i=1,\ldots,k$.

The Riemannian polynomial equations cannot, in general, be solved in closed form, and must be integrated numerically. In order to discretize this system of covariant differential equations, we implement a covariant Euler integrator, depicted in Algorithm 1. A time step Δt is chosen and, at each step of the integrator, $\gamma(t+\Delta t)$ is computed using the exponential map:

$$\gamma(t+\Delta t)=\mathrm{Exp}_{\gamma(t)}\big(\Delta t\, v_1(t)\big). \tag{18}$$

Each vector v_i is incremented within the tangent space at $\gamma(t)$ and the results are parallel transported infinitesimally along a geodesic from $\gamma(t)$ to $\gamma(t+\Delta t)$. For a proof that this algorithm approximates the polynomial equations, see Appendix A. The only ingredients necessary to integrate a polynomial are the exponential map and parallel transport on the manifold.

Algorithm 1 Pseudocode for forward integration of k^{th} order Riemannian polynomial

```
γ ← γ(0)
for i = 1, ..., k do
    v_i ← v_i(0)
end for
t ← 0
repeat
    w ← v_1
    for i = 1, ..., k − 1 do
        v_i ← ParTrans(γ, Δt w, v_i + Δt v_{i+1})
    end for
    v_k ← ParTrans(γ, Δt w, v_k)
    γ ← Exp_γ(Δt w)
    t ← t + Δt
until t=T
```

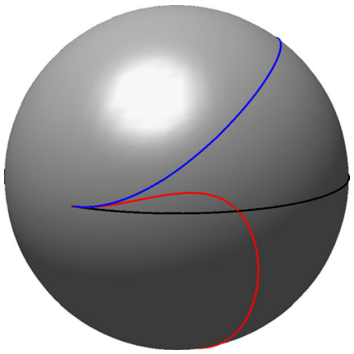


Fig. 1 Sample polynomial curves emanating from a common basepoint on the sphere (*black* = geodesic, *blue* = quadratic, *red* = cubic)

Figure 1 shows the result of integrating polynomials of order one, two, and three on the sphere. The parameters, the initial velocity, acceleration, and jerk, were chosen a priori and a cubic polynomial was integrated to obtain the blue curve. Then the initial jerk was set to zero and the blue quadratic curve was integrated, followed by the black geodesic whose acceleration was also set to zero.

3.1 Polynomial Time Reparametrization

Geodesic curves propagate at a constant speed as a result of their extremal action property. Polynomials provide flexibility not only in the class of paths that are possible, but in the time dependence of the curves traversing those paths. If the parameters of a polynomial γ consist of collinear vectors

$v_i(0) \in T_{\gamma(0)}M$, then the path of γ (the image of the mapping γ) matches that of a geodesic, but the time dependence has been reparametrized by some polynomial transformation $t \mapsto c_0 + c_1 t + c_2 t^2 + c_3 t^3$. This generalizes the existence of polynomials in Euclidean space which are merely polynomial transformations of a straight line path. Regression models could even be implemented in which the operator wishes to estimate geodesic paths, but is unsure of parametrization, and so enforces the estimated parameters to be collinear.

4 Polynomial Regression via Adjoint Optimization

In order to regress polynomials against observed data $J_j \in M$, $j = 1, \ldots, N$ at known times $t_j \in \mathbb{R}$, $j = 1, \ldots, N$, we define the following objective function

$$E_0\big(\gamma(0), v_1(0), \ldots, v_k(0)\big) = \frac{1}{N}\sum_{j=1}^{N} d\big(\gamma(t_j), J_j\big)^2 \tag{19}$$

subject to the constraints given by Eqs. (15)–(17). Note that in this expression d represents the geodesic distance: the minimum length of a path from the curve point $\gamma(t_j)$ to the data point J_j. The function E_0 is minimized in order to find the optimal initial conditions $\gamma(0), v_i(0), i = 1, \ldots, k$, which we will refer to as the parameters of our model.

In order to determine the optimal parameters of the polynomial, we introduce Lagrange multiplier vector fields λ_i for $i = 0, \ldots, k$, often called the adjoint variables, and define the augmented Lagrangian function

$$\begin{aligned} &E\big(\gamma, \{v_i\}, \{\lambda_i\}\big) \\ &\quad= \frac{1}{N}\sum_{j=1}^{N} d\big(\gamma(t_j), J_j\big)^2 \\ &\qquad+ \int_0^T \left\langle \lambda_0(t), \frac{d}{dt}\gamma(t) - v_1(t) \right\rangle dt \\ &\qquad+ \sum_{i=1}^{k-1} \int_0^T \big\langle \lambda_i(t), \nabla_{\frac{d}{dt}\gamma} v_i(t) - v_{i+1}(t) \big\rangle dt \\ &\qquad+ \int_0^T \big\langle \lambda_k(t), \nabla_{\frac{d}{dt}\gamma} v_k(t) \big\rangle dt. \end{aligned} \tag{20}$$

As is standard practice, the optimality conditions for this equation are obtained by taking variations with respect to all arguments of E, integrating by parts when necessary. The resulting variations with respect to the adjoint variables yield the original dynamic constraints: the polynomial equations. Variations with respect to the primal variables gives rise to the following system of equations, termed the adjoint equations (see B for derivation).

$$\nabla_{\frac{d}{dt}\gamma}\lambda_i(t) = -\lambda_{i-1}(t) \quad i = 1, \ldots, k \tag{21}$$

$$\nabla_{\frac{d}{dt}\gamma}\lambda_0(t) = -\sum_{i=1}^{k} R\big(v_i(t), \lambda_i(t)\big)v_1(t), \tag{22}$$

where R is the Riemannian curvature tensor and the adjoint variable λ_0 takes jump discontinuities at time points where data is present:

$$\lambda_0\big(t_j^-\big) - \lambda_0\big(t_j^+\big) = \mathrm{Log}_{\gamma(t_j)} J_j. \tag{23}$$

Note that this jump discontinuity corresponds to the variation of E with respect to $\gamma(t_j)$. The Riemannian curvature tensor is defined by the formula [9]

$$R(u, v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w, \tag{24}$$

and can be computed in closed form for many manifolds. Gradients of E with respect to initial and final conditions give rise to the terminal endpoint conditions for the adjoint variables,

$$\lambda_i(1) = 0, \quad i = 0, \ldots, k \tag{25}$$

as well as expressions for the gradients with respect to the parameters $\gamma(0), v_i(0)$:

$$\delta_{\gamma(0)} E = -\lambda_0(0), \tag{26}$$

$$\delta_{v_i(0)} E = -\lambda_i(0). \tag{27}$$

In order to determine the value of the adjoint vector fields at $t = 0$, and thus the gradients of the functional E_0, the adjoint variables are initialized to zero at time 1, then Eq. (22) is integrated backward in time to $t = 0$.

Given the gradients with respect to the parameters, a simple steepest descent algorithm is used to optimize the functional. At each iteration, $\gamma(0)$ is updated using the exponential map and the vectors $v_i(0)$ are updated via parallel translation. This algorithm is depicted in Algorithm 2.

Note that in the special case of a zero-order polynomial ($k = 0$), the only gradient λ_0 is simply the mean of the log map vectors at the current estimate of the Fréchet mean. So this method generalizes the common method of Fréchet averaging on manifolds via gradient descent [13]. In the case of geodesic polynomials, $k = 1$, the curvature term in Eq. (22) indicates that λ_1 is a sum of Jacobi fields. So this approach subsumes geodesic regression as presented by Fletcher [12]. For higher order polynomials, the adjoint equations represent a generalization of Jacobi field.

As we will see later, in some cases these adjoint equations take a simpler form not involving curvature. In the case that the manifold M is a Lie group, the adjoint equations can be computed by taking variations in the Lie algebra, avoiding explicit curvature computation.

Algorithm 2 Pseudocode for reverse integration of adjoint equations for k^{th} order Riemannian polynomial

$\gamma \leftarrow \gamma(T)$
for $i = 0, \ldots, k$ **do**
  $\lambda_i \leftarrow 0$
end for
$t \leftarrow T$
repeat
  $w \leftarrow v_1(t)$
  $\lambda_0 \leftarrow \lambda_0 + \Delta t \sum_{i=1}^{k} R(v_i, \lambda_i) v_1$
  if $t = t_i$ **then**
    $\lambda_0 \leftarrow \lambda_0 + \frac{2}{N} \operatorname{Log}_\gamma J_i$
  end if
  for $i = k, \ldots, 1$ **do**
    $\lambda_i \leftarrow \operatorname{ParTrans}(\gamma, -\Delta t w, \lambda_i + \Delta t \lambda_{i-1})$
  end for
  $\lambda_0 \leftarrow \operatorname{ParTrans}(\gamma, -\Delta t w, \lambda_0)$
  $\gamma \leftarrow \operatorname{Exp}_\gamma(-\Delta t w)$
  $t \leftarrow t - \Delta t$
until t=0
$\delta_{\gamma(0)} E \leftarrow -\lambda_0$
for $i = 1, \ldots, k$ **do**
  $\delta_{v_i(0)} E \leftarrow -\lambda_i$
end for

4.1 Coefficient of Determination (R^2) in Metric Spaces

In order to characterize how well our model fits a given set of data, we define the coefficient of determination of our regression curve $\gamma(t)$, denoted R^2 [12]. As with the usual definition of R^2, we first compute the variance of the data. Naturally, as the data lie on a non-Euclidean metric space, instead of the standard sample variance, we substitute the Fréchet variance, defined as

$$\operatorname{var}\{y_1, \ldots, y_N\} = \frac{1}{N} \min_{\bar{y} \in M} \sum_{j=1}^{N} d(\bar{y}, y_j)^2. \tag{28}$$

The sum of squared error for a curve γ is the value $E_0(\gamma)$:

$$SSE = \frac{1}{N} \sum_{j=1}^{N} d\big(\gamma(t_j), y_j\big)^2. \tag{29}$$

We then define R^2 as the amount of variance that has been reduced using the curve γ:

$$R^2 = 1 - \frac{SSE}{\operatorname{var}\{y_1, \ldots, N\}}. \tag{30}$$

Clearly a perfect fit will remove all error, resulting in an R^2 value of one. The worst case ($R^2 = 0$) occurs when no polynomial can improve over a stationary point at the Fréchet mean, which can be considered a zero-order polynomial regression against the data.

4.2 Example: Kendall Shape Space

A common challenge in medical imaging is the comparison of shape features which are independent of easily explained differences such as differences in pose (relative position and rotation). Additionally, scale is often uninteresting as it is easily characterized by volume calculation and explained mostly by intersubject variability or differences in age. It was with this perspective that Kendall [19] originally developed his theory of shape space. Here we briefly describe Kendall's shape space of m-landmark point sets in $\mathbb{R}^d$, denoted Σ_d^m. For a complete treatment of Kendall's shape space, the reader is encouraged to consult Kendall and Le [20, 23].

Given a point set $x = (x_i)_{i=1,\ldots,m}, x_i \in \mathbb{R}^d$, translation and scaling effects are removed by centering and uniform scaling. This is achieved by translating the point set so that the centroid is at zero, then scaling so that $\sum_{i=1}^{m} \|x_i\|^2 = 1$. After this standardization, x constitutes a point in the sphere $\mathbb{S}^{(m-1)d-1}$. This representation of shape is not yet complete as it is effected by global rotation, which we wish to ignore. Thus points on $\mathbb{S}^{(m-1)d-1}$ are referred to as *preshapes* and the sphere $\mathbb{S}^{(m-1)d-1}$ is referred to as *preshape space*. Kendall shape space Σ_d^m is obtained by taking the quotient of the preshape space by the action of the rotation group $\mathrm{SO}(d)$. In practice, points in the quotient (referred to as *shapes*) are represented by members of their equivalence class in preshape space. We describe now how to compute exponential maps, log maps, and parallel transport in shape space, using representatives in $\mathbb{S}^{(m-1)d-1}$. The work of O'Neill [33] concerning Riemannian submersions characterizes the link between the shape and preshape spaces.

The case $d > 2$ is complicated in that these spaces contain degeneracies: points at which the mapping from preshape space to Σ_d^m fails to be a submersion [1, 11, 17]. Despite these pathologies, outside of a singular set, the shape spaces are described by the theory of Riemannian submersions. We assume the data lie within a single "manifold part" away from any singularities, and show experiments in two dimensions, so that these technical issues can be safely ignored.

Each point p in preshape space projects to a point $\pi(p)$ in shape space. The shape $\pi(p)$ is the orbit of p under the action of $\mathrm{SO}(d)$. Viewed as a subset of $\mathbb{S}^{(m-1)d-1}$, this orbit is a submanifold whose tangent space is a subspace of that of the sphere. This subspace is called the vertical subspace of $T_p\mathbb{S}^{(m-1)d-1}$ and its orthogonal complement is the horizontal subspace. Projections onto the two subspaces of a vector $v \in T_p\mathbb{S}^{(m-1)d-1}$ are denoted by $\mathcal{V}(v)$ and $\mathcal{H}(v)$, respectively. Curves moving along vertical tangent vectors result in rotations of a preshape, and so do not indicate any change in actual shape.

A vertical vector in preshape space arises as the derivative of a rotation of a preshape. The derivative of such a rotation

is a skew-symmetric matrix W, and its action on a preshape x has the form $(Wx_1, \ldots, Wx_n) \in T\mathbb{S}^{(m-1)d-1}$. The vertical subspace is then spanned by such tangent vectors arising from any linearly independent set of skew-symmetric matrices. The projection $\mathcal{H}$ is performed by taking such a spanning set, performing Gram-Schmidt orthonormalization, and removing each component.

The horizontal projection allows one to relate the covariant derivative on the sphere to that on shape space. Lemma 1 of O'Neill [33] states that if X, Y are horizontal vector fields at some point p in preshape space, then

$$\mathcal{H}\nabla_X Y = \nabla^*_{X^*} Y^*, \tag{31}$$

where ∇ denotes the covariant derivative on preshape space and ∇^*, X^*, and Y^* are their counterparts in shape space.

For the manifold part of a general shape space Σ_d^m, the exponential map and parallel translation are performed using representatives preshapes in $\mathbb{S}^{(m-1)d-1}$. For $d > 2$, this must be done in a time-stepping algorithm, in which at each time step an infinitesimal spherical parallel transport is performed, followed by the horizontal projection. The resulting algorithm can be used to compute the exponential map as well. Computation of the log map is less trivial, as it requires an iterative optimization routine. A special case arises in the case when $d = 2$, in which case the entire space Σ_d^m is a manifold. In this case the exponential map, parallel transport and log map are computed in closed form [12]. With the exponential map, log map, and parallel transport, one performs polynomial regression on Kendall shape space via the adjoint method described previously.

4.2.1 Rat Calivaria Growth

We have applied polynomial regression in Kendall shape space to the data first analyzed by Bookstein [2], which consists of $m = 8$ landmarks on a midsagittal section of rat calivaria (skulls excluding the lower jaw). The positions of eight identifiable positions on the skull are available for 18 rats and at of eight ages apiece. Figure 2 shows Riemannian polynomial fits of orders $k = 0, 1, 2, 3$. Curves of the same color indicate the synchronized motion of landmarks within a preshape, and the collection of curves for all eight landmarks represents a curve in shape space. While the geodesic curve in Kendall shape space shows little curvature, the quadratic and cubic curves are less linear which demonstrates the added flexibility provided by higher order polynomials. The R^2 values agree with this qualitative difference: the geodesic regression has $R^2 = 0.79$, while the quadratic and cubic regressions have R^2 values of 0.85 and 0.87, respectively. While this shows that there is a clear improvement in the fit due to increasing k from one to two, it also shows that little is gained by increasing the order of the polynomial beyond $k = 2$. Qualitatively, Fig. 2 shows that the slight increase in R^2 obtained by moving from a quadratic to cubic model corresponds to a marked difference in the curves, indicating that the cubic curve is likely overfitting the data. As seen in Table 1, increasing the order of polynomial to four or five has very little effect on R^2 as well.

These results indicate that moving from a geodesic to quadratic model provides an important improvement in fit quality. This is consistent with the results of Kenobi et al. [21], who also found that quadratic and possibly cubic curves are necessary to fit this dataset. However, whereas Kenobi et al. use polynomials defined in the tangent space

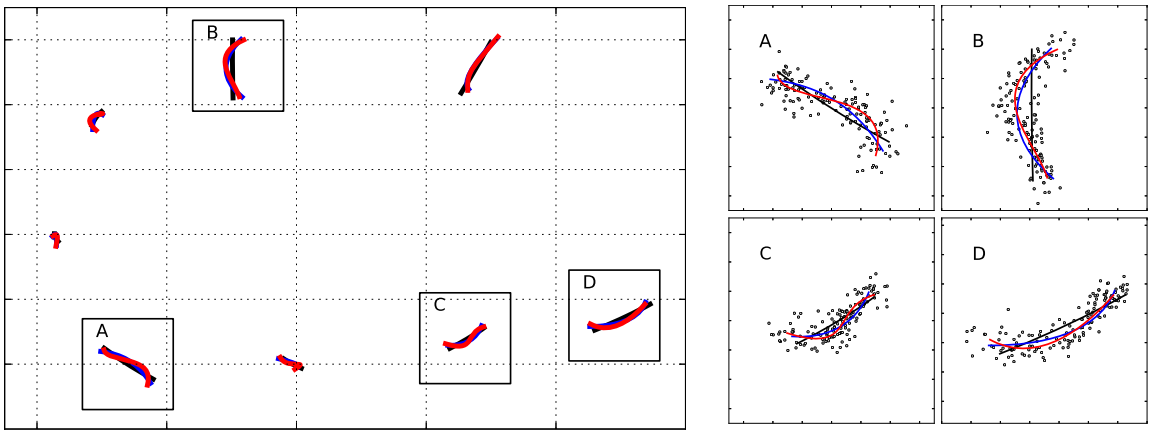


Fig. 2 Bookstein rat calivaria data after uniform scaling and Procrustes alignment. The colors of lines indicate order of polynomial used for the regression (*black* = geodesic, *blue* = quadratic, *red* = cubic). Zoomed views of individual rectangles are shown at right, along with data points in *gray*. Note that the axes are arbitrary, due to scale-invariance of Kendall shape space, but that they are the same for the horizontal and vertical axes in these figures

Table 1 R^2 for regression of rat dataset

Polynomial order k	R^2
1	0.79
2	0.85
3	0.87
4	0.87
5	0.87

Table 2 R^2 for regression of corpus callosum dataset

Polynomial order k	R^2
1	0.11
2	0.14
3	0.20
4	0.21
5	0.22

at the Fréchet mean of the data points, the polynomials we use are defined intrinsically, independent of base point.

4.2.2 *Corpus Callosum Aging*

The corpus callosum, the major white matter bundle connecting the two hemispheres of the brain, is known to shrink during aging [10]. Fletcher showed [12] that more nuanced modes of shape change are observed using geodesic regression. In particular, the volume change observed in earlier studies corresponds to a thinning of the corpus callosum and increased curling of the anterior and posterior regions. In order to investigate even higher modes of shape change of the corpus callosum during normal aging, polynomial regression was performed on data from the OASIS brain database [27]. Magnetic resonance imaging (MRI) scans from 32 normal subjects with ages between 19 and 90 years were obtained from the database and a midsagittal slice was extracted from each volumetric image. The corpus callosum was then segmented on the 2D slices using the ITK-SNAP program [39]. Sets of 64 landmarks for each patient were obtained using the ShapeWorks program [6], which generates samplings of each shape boundary with optimal correspondences among the population.

Regression results for geodesic, quadratic, and cubic regression are shown in Fig. 3. At first glance the results appear similar for the three different models, since the motion envelopes each show the thinning and curling observed by Fletcher. Indeed, the optimal quadratic curve is quite similar to the optimal geodesic, as reflected by their similar R^2 values (0.13 and 0.12, respectively). However, moving from a quadratic to cubic polynomial model delivers a substantial increase in R^2 (from 0.13 to 0.21). This suggests that there are interesting third-order phenomena at work. However, as seen in Table 2, increasing the order beyond three results in very little increase in R^2, indicating that those orders overfit the data, as was the case in the rat calivaria study as well.

Inspection of the estimated parameters for the optimal cubic curve, shown in Fig. 4, reveals that the tangent vectors appear to be collinear. As discussed in Sect. 3.1, this suggests that the cubic curve is a geodesic that has undergone a cubic time reparametrization.

Note that the R^2 values are quite low in this study. Similar values were observed using geodesic regression in [12].

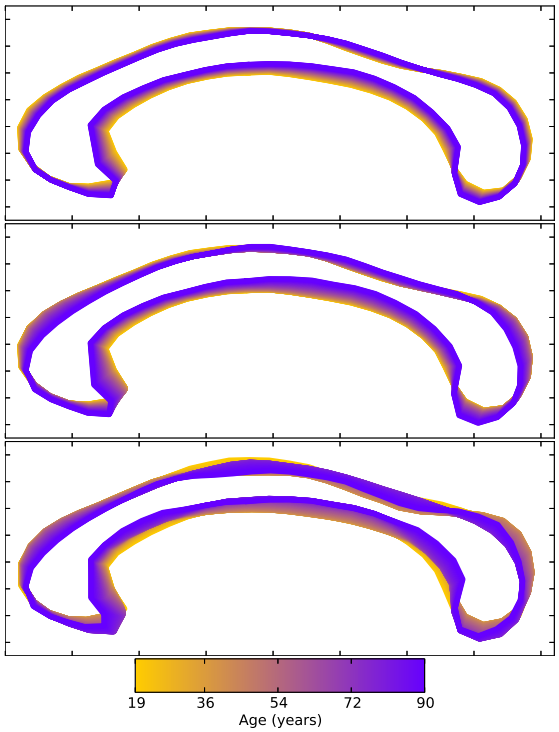


Fig. 3 Geodesic (*top*, $R^2 = 0.12$) quadratic (*middle*, $R^2 = 0.13$) and cubic (*bottom*, $R^2 = 0.21$) regression for corpus callosum dataset. Color represents age, with *yellow* indicating youth (age 19) and *purple* indicating old age (age 90)

As is noted, this is likely due to high inter-subject variability, and that age is only able to explain an effect which is small compared to differences between subjects. Fletcher [12] also notes that although the effect may be small, geodesic regression gives a result which is significant ($p = 0.009$) using a non-parametric permutation test.

Model selection, which in the case of polynomial regression amounts to the choice of polynomial order, is an important issue. R^2 always increases with increasing k, as we have seen in these two studies. As a result, other measures are sought which balance goodness of fit with complexity of the curve model. Tools often used for model selection in Euclidean polynomial regression, such as Akaike informa-

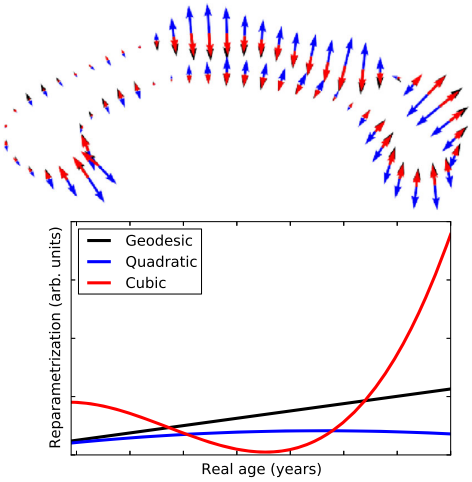


Fig. 4 Parameters for regression of corpus callosa using a cubic polynomial. The velocity (*black*), acceleration (*blue*) and jerk (*red*) are nearly collinear, indicating that the estimated path is essentially a geodesic with cubic time reparametrization. The time reparametrization is shown in the plot, for geodesic, quadratic, and cubic Riemannian polynomial regression

tion criterion and Bayesian information criterion [5] make assumptions about the distribution of data that are difficult to generalize to the manifold setting. Extension of permutation testing for geodesic regression to higher orders would be useful for this task, but such extension is not trivial on a Riemannian manifold. We expect that such an extension of permutation testing is possible in certain cases where it is possible to define "exchangeability" under the null hypothesis that the data follow a given order k trend. Currently, we select models based on qualitative analysis of the fit curves, as in the rat calivaria study, and R^2 values.

4.3 LDDMM Landmark Space

Analysis of landmarks is commonly done in an alternative fashion when scale and rotation invariance is not desired. In this section, we present polynomial regression using the large distance diffeomorphic metric mapping (LDDMM) framework. This framework consists of a Lie group of diffeomorphisms endowed with a right invariant Sobolev metric acting on a space of landmark configurations. For a more detailed description of the group action approach, the reader is encouraged to consult Bruveris et al. [4]. We will instead focus on the Riemannian structure of landmarks and use the formulas for general Riemannian manifolds.

Given m landmarks in d dimensions, let $M \cong \mathbb{R}^{md}$ be the space of all possible configurations. We denote by $x_i \in \mathbb{R}^d$ the location of the ith landmark point. Tangent vectors are also represented as tuples of vectors, $v = (v_i)_{i=1,\dots,m} \in \mathbb{R}^{md}$, as are cotangent vectors $\alpha = (\alpha_i)_{i=1,\dots,m} \in \mathbb{R}^{md}$. Contrasting ordinary differential geometric methods in which vectors and metrics are the objects of interest, it is more convenient to work with landmark covectors (which we refer to as momenta). In such case the inverse metric (also called the cometric) is generally written using a shift-invariant scalar kernel $K : \mathbb{R} \to \mathbb{R}$. The inner product of two covectors is given by

$$\langle \alpha, \beta \rangle_{T_x^* M} = \sum_{i,j} K\big(|x_i - x_j|^2\big)\alpha_i^T \beta_j. \tag{32}$$

The following Hamilton's equations describe geodesics in landmark space [38, Eq. (21)]:

$$\frac{d}{dt}x_i = \sum_j K\big(|x_i - x_j|^2\big)\alpha_j \tag{33}$$

$$\frac{d}{dt}\alpha_i = \sum_j 2(x_i - x_j)K'\big(|x_i - x_j|^2\big)\alpha_i^T \alpha_j \tag{34}$$

where K' denotes the derivative of the kernel.

Introducing tangent vectors $v = K\alpha$ and $w = K\beta$, parallel transport in LDDMM landmark space are computed in coordinates using the following formula, derived by Younes et al. [38, Eq. (25)]:

$$\begin{aligned}\frac{d}{dt}\beta_i = K^{-1}\Bigg(&\sum_{j=1}^{N}(x_i - x_j)^T(w_i - w_j)K'\big(|x_i - x_j|^2\big)\alpha_j \\ &- \sum_{j=1}^{N}(x_i - x_j)^T(v_i - v_j)K'\big(|x_i - x_j|^2\big)\beta_j\Bigg) \\ &- \sum_{j=1}^{N}(x_i - x_j)\gamma'\big(|x_i - x_j|^2\big)\big(\alpha_j^T\beta_i + \alpha_i^T\beta_j\big).\end{aligned} \tag{35}$$

In order to integrate the adjoint equations, it is also necessary to compute the Riemannian curvature tensor, which in this case is more complicated. For an in-depth treatment, see Micheli et al. [29, Theorem 2.2].

Using these approaches to computing parallel transport and curvature, we implemented the general polynomial adjoint optimization method. We applied this approach to the rat calivaria data, treating the data as absolute landmark positions (after Procrustes alignment) instead of as scale and rotation invariant Kendall shapes.

Shown in Fig. 5 are the results of LDDMM landmark polynomial regression. Notice that while the geodesic curve in this case corresponds to nonlinear trajectories for the individual landmarks, these paths do not fit the data quite as well as the quadratic curve. In particular, the point at the crown of the skull (labelled point A in Fig. 5) appears to change directions in the quadratic curve, which is not possible using

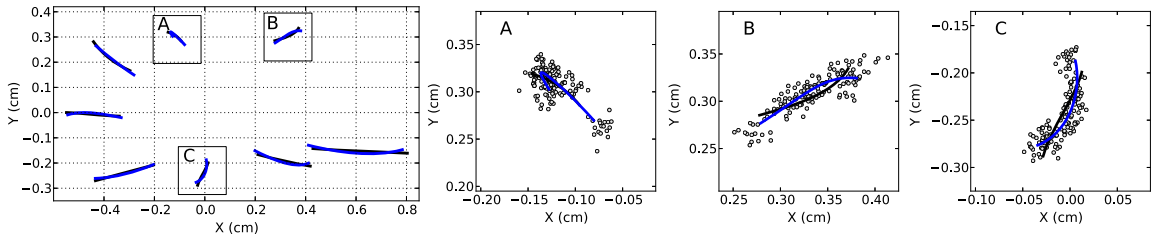


Fig. 5 Regression curves for Bookstein rat data using LDDMM landmark polynomials. The colors of lines indicate order of polynomial used for the regression (*black* = geodesic, *blue* = quadratic). Zoomed views of individual rectangles are shown at right, along with data points in *gray*. The data were aligned with respect to translation and rotation but not scaling, which explains the clear growth trend

a geodesic. These qualitative improvements correspond to a slight increase in R^2, from 0.92 with the geodesic to 0.94 with the quadratic curve.

5 Riemannian Polynomials in Lie Groups

In this section, we consider the case when the configuration manifold is a Lie group G. A tangent vector $v \in T_gG$ at a point $g \in G$ can be identified with a tangent vector at the identity element $e \in G$ via either right or left translation by g^{-1}. The resulting element of T_eG is referred to as the right (respectively, left) trivialization of v. We call a vector field $X \in \mathfrak{X}(G)$ right (respectively, left) invariant if the right trivialization of $X(g)$ is constant for all g. Both left and right translation, considered as mappings $T_gG \to T_eG$ are linear isomorphisms, and we will use the common notation $\mathfrak{g}$ to refer to T_eG. The vector space $\mathfrak{g}$, endowed with the vector product given by the right trivialization of the negative Jacobi-Lie bracket of right invariant vector fields is called the Lie algebra of G.

Of particular importance to the study of Lie groups is the adjoint representation, which for each group element g determines a linear action Ad_g on $\mathfrak{g}$ called the adjoint action and its dual action Ad^*_g on $\mathfrak{g}^*$ which is called the coadjoint action of g. In a Riemannian Lie group, the inner product on $\mathfrak{g}$ can be used to compute the adjoint of the adjoint action, which we term the adjoint-transpose action $\mathrm{Ad}^\dagger_g$, defined by

$$\left\langle \mathrm{Ad}^\dagger_g X, Y \right\rangle = \langle X, \mathrm{Ad}_g Y \rangle \tag{36}$$

for all $X, Y \in \mathfrak{g}$. The infinitesimal version of these actions at the identity element are termed the infinitesimal adjoint action, ad_X, and the infinitesimal adjoint-transpose action, $\mathrm{ad}^\dagger_X$. These operators, along with the metric at the identity, encode all geometric properties such as covariant derivatives and curvature in a Lie group with right invariant Riemannian metric. For a more complete review of Lie groups and the adjoint representation, see [28]. Following [25], we introduce the symmetric product of two vectors $X, Y \in \mathfrak{g}$ as

$$\mathrm{sym}_X Y = \mathrm{sym}_Y X = -\left(\mathrm{ad}^\dagger_X Y + \mathrm{ad}^\dagger_Y X\right). \tag{37}$$

Extending X and Y to right invariant vector fields $\widetilde{X}, \widetilde{Y}$, the covariant derivative $\nabla_{\widetilde{X}}\widetilde{Y}$ is also right invariant (c.f. [7, Proposition 3.18]) and satisfies

$$(\nabla_{\widetilde{X}}\widetilde{Y})g^{-1} = -\overline{\nabla}_X Y \tag{38}$$

where we have introduced the notation $\overline{\nabla}$ for the reduced Levi-Civita connection:

$$\overline{\nabla}_X Y = \frac{1}{2}\,\mathrm{ad}_X Y + \frac{1}{2}\,\mathrm{sym}_X Y. \tag{39}$$

Notice that in this notation, ad represents the skew-symmetric component of the Levi-Civita connection, while sym represents the symmetric component.

We use ξ_1 to denote the right trivialized velocity of the curve $\gamma(t) \in G$. Using our formula for the covariant derivative, one sees that the geodesic equation in a Lie group with right invariant metric is the right "Euler-Poincaré" equation:

$$\frac{d}{dt}\xi_1 = \overline{\nabla}_{\xi_1}\xi_1 = -\,\mathrm{ad}^\dagger_{\xi_1}\,\xi_1. \tag{40}$$

The left Euler-Poincaré equation is obtained by removing the negative sign from the right hand side. For polynomials, the Euler-Poincaré equation is generalized to higher order. Introducing $\xi_i, i = 1, \ldots, k$ to represent the right trivialized higher-order velocity vectors v_i,

$$v_i(t) = \xi_i(t)g(t), \tag{41}$$

the reduced Riemannian polynomial equations are

$$\frac{d}{dt}\gamma(t) = \xi_1\gamma(t) \tag{42}$$

$$\frac{d}{dt}\xi_i(t) = \overline{\nabla}_{\xi_1}\xi_i(t) + \xi_{i+1}(t), \quad i = 1, \ldots, k-1 \tag{43}$$

$$\frac{d}{dt}\xi_k(t) = \overline{\nabla}_{\xi_1}\xi_k(t). \tag{44}$$

Notice that these equations correspond precisely to the polynomial equations (Eq. (15)).

6 Polynomial Regression in Lie Groups

We have seen that the geodesic equation is simplified in a Lie group with right invariant metric, using the Euler-Poincaré equation. In this section, we derive the adjoint equations used to perform geodesic and polynomial regression in a Lie group. Using right-trivialized adjoint variables, we will see that the symmetries provided by the Lie group structure result in adjoint equations more amenable to computation than those in Sect. 4.

6.1 Geodesic Regression

Before moving on to polynomial regression, we first present an adjoint optimization approach to geodesic regression in a Lie group with right invariant metric. Suppose N data points $J_j \in G$ are observed at times $t_j \in [0,1]$. Using the geodesic distance $d : G \times G \to \mathbb{R}$, the least squares geodesic regression problem is to find the minimum of

$$E(\gamma) = \frac{1}{2}\sum_{j=1}^{N} d\big(\gamma(t_j), J_j\big)^2, \tag{45}$$

subject to the constraint that the curve $\gamma : [0,1] \to G$ is a geodesic.

In order to determine optimality conditions for γ, consider a variation of the geodesic $\gamma(t)$, which is a vector field along γ that we denote $\delta\gamma(t) \in T_{\gamma(t)}G$. We denote by $Z(t)$ the right trivialization of $\delta\gamma(t)$. The variation of γ induces the following variation in the trivialized velocity ξ_1 [16]:

$$\delta\xi_1(t) = \frac{d}{dt}Z(t) - \mathrm{ad}_{\xi_1} Z(t). \tag{46}$$

Constraining $\delta\gamma$ to be a Jacobi field, we use the following variation of the Euler-Poincaré equation to obtain

$$\frac{d}{dt}\delta\xi_1 = \delta\left(\frac{d}{dt}\xi_1\right) = \delta\big(-\mathrm{ad}^{\dagger}_{\xi_1}\xi_1\big) = \mathrm{sym}_{\xi_1}\delta\xi_1. \tag{47}$$

Combining these results, we write the ordinary differential equation (ODE) that determines, along with initial conditions, the vector field Z:

$$\frac{d}{dt}\begin{pmatrix} Z \\ \delta\xi_1 \end{pmatrix} = \begin{pmatrix} \mathrm{ad}_{\xi_1} & I \\ 0 & \mathrm{sym}_{\xi_1} \end{pmatrix}\begin{pmatrix} Z \\ \delta\xi_1 \end{pmatrix}. \tag{48}$$

This ODE constitutes a general perturbation of a geodesic and the vector field $Z(t)$ is a right trivialized Jacobi field. In order to compute the variations of E with respect to the initial position $\gamma(0)$ and velocity $\xi_1(0)$ of the geodesic $\gamma(t)$, the variations of E with respect to $\gamma(1)$ and $\xi_1(1)$ are transported backward to $t = 0$ by the adjoint ODE. Introducing adjoint variables $\lambda_0(t), \lambda_1(t) \in \mathfrak{g}$, the left trivialized variation of E with respect to $\gamma(t)$ and the variation with respect to $\xi_1(t)$ are given by

$$\delta_{\gamma(0)}E = -\lambda_0(0) \tag{49}$$

$$\delta_{\xi_1(0)}E = -\lambda_1(0). \tag{50}$$

These variations are computed by initializing $\lambda_0(1) = \lambda_1(1) = 0$ and integrating the adjoint ODE backward to $t = 0$. The adjoint ODE is obtained by simply computing the adjoint of the ODE governing geodesic perturbations, Eq. (48), with respect to the $L^2([0,1] \to \mathfrak{g})$ inner product. The resulting adjoint ODE is

$$\frac{d}{dt}\begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix} = \begin{pmatrix} -\mathrm{ad}^{\dagger}_{\xi_1} & 0 \\ -I & -\mathrm{sym}^{\dagger}_{\xi_1} \end{pmatrix}\begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix}, \tag{51}$$

where the adjoint of the symmetric product is given by

$$\mathrm{sym}^{\dagger}_X Y = -\mathrm{ad}_X Y + \mathrm{ad}^{\dagger}_Y X. \tag{52}$$

The adjoint variable λ_0 takes jump discontinuities when passing over data points:

$$\lambda_0\big(t_j^-\big) - \lambda_0\big(t_j^+\big) = (\mathrm{Log}_{\gamma(t_j)} J_j)\gamma(t_j)^{-1}. \tag{53}$$

The jumps represent the residual vectors, obtained by right trivialization of the Riemannian log map from the predicted point $\gamma(t_j)$ to the data J_j. Notice that the adjoint variable λ satisfies an equation resembling the Euler-Poincaré equation and can likewise be solved in closed form:

$$\lambda_0(t) = \sum_{j,t_j>t} \mathrm{Ad}^{\dagger}_{\gamma^{-1}(t)\gamma(t_j)}\,\mathrm{Log}_{\gamma(t_j)} J_j. \tag{54}$$

This is particularly useful because it reduces the second order ODE, Eq. (51), to an ODE of first order, since the first equation is solved in closed form. We will soon see that this simplification occurs even when using higher order polynomials.

Finally, minimization of E is performed using the variations $\delta_{\gamma(0)}E, \delta_{\xi_1(0)}E$ using, for example the following gradient descent steps:

$$\gamma(0)^{k+1} = \mathrm{Exp}(-\alpha\delta_{\gamma(0)^k}E)\gamma(0)^k \tag{55}$$

$$\xi_1(0)^{k+1} = \xi_1(0)^k - \alpha\delta_{\xi(0)^k}E \tag{56}$$

for some positive step size α, where k denotes the step of the iterative optimization process. Note that commonly the Riemannian exponential map Exp in the above expression is replaced by a numerically efficient approximation such as the Cayley map [3].

6.2 Example: Rotation Group SO(3)

As an example, in this section we derive the algorithm for polynomial regression in the group of rotations in three dimensions, SO(3). This group consists of orthogonal matrices with determinant one, and has associated the Lie algebra $\mathfrak{so}(3)$ of skew-symmetric 3-by-3 matrices. Skew-symmetric matrices can be bijectively identified with vectors in $\mathbb{R}^3$ using the following mapping $*$:

$$* : \mathbb{R}^3 \leftrightarrow \mathfrak{so}(3), \qquad * \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}. \tag{57}$$

We use a star to indicate both this mapping $\mathbb{R}^3 \to \mathfrak{so}(3)$ and its inverse, a notation which emphasizes that it is the Hodge dual in $\mathbb{R}^3$, though it is also commonly written using a hat symbol [28]. Using the cross product on $\mathbb{R}^3$, the star map is also a Lie algebra isomorphism, so that

$$* \operatorname{ad}_{*x} * y = x \times y. \tag{58}$$

The adjoint action under the star is also quite convenient, as it is given simply by matrix-vector multiplication:

$$\operatorname{Ad}_g(*x) = *(gx) \tag{59}$$

for any $g \in \mathrm{SO}(3)$, $x \in \mathbb{R}^3$.

We will use a left invariant metric given by a symmetric positive definite 3-by-3 matrix A. For vectors $x, y \in \mathbb{R}^3$, the inner product is

$$\langle *x, *y \rangle_{\mathfrak{g}} = x^T A y. \tag{60}$$

With this inner product, the infinitesimal adjoint transpose action is

$$* \operatorname{ad}^\dagger_{*x} * y = -A^{-1}(x \times Ay). \tag{61}$$

The most natural metric is that in which A is the identity matrix. In that case, left invariance also implies right invariance and skew-symmetry of $\operatorname{ad}^\dagger$, so that for any $X, Y \in \mathfrak{so}(3)$:

$$\operatorname{sym}_X Y 0, \qquad \overline{\nabla}_X Y = \frac{1}{2} \operatorname{ad}_X Y. \tag{62}$$

The Euler-Poincaré equation in the biinvariant case is

$$\frac{d}{dt}\xi = \operatorname{ad}^\dagger_\xi \xi = - * \xi \times * \xi = 0, \tag{63}$$

implying that geodesics using the biinvariant metric have constant trivialized velocity. The geodesic can then be integrated in closed form:

$$\frac{d}{dt}\gamma(t) = \xi \gamma(t) \quad \Longrightarrow \quad \gamma(t) = \exp(t\xi). \tag{64}$$

Notice that the adjoint-transpose action of a rotation matrix $g \in \mathrm{SO}(3)$ on a 3-vector x is given by

$$* \operatorname{Ad}^\dagger_g(*x) = g^T x. \tag{65}$$

So the first adjoint equation is given by

$$\begin{aligned} \lambda_0(t) &= \gamma(t)^T \gamma(1) \lambda_0(1) && (66)\\ &= \exp(-t\xi)\exp(\xi)\lambda_0(1) && (67)\\ &= \exp\big((1-t)\xi\big)\lambda_0(1) && (68)\\ &= \lambda_0(1)\cos\big((1-t)\|\xi\|\big) \\ &\quad - \frac{1}{\|\xi\|}\big(*\xi \times \lambda_0(1)\big)\sin\big((1-t)\|\xi\|\big) \\ &\quad + \frac{1}{\|\xi\|^2} * \xi\big(*\xi \cdot \lambda_0(1)\big)\big(1 - \cos\big((1-t)\|\xi\|\big)\big). && (69) \end{aligned}$$

where the last line is Rodrigues' rotation formula. The second adjoint equation, which determines the variation used to update the velocity, is obtained by integrating this. For geodesic regression with biinvariant metric, a closed form solution is available for the second adjoint variable as well:

$$\begin{aligned} \frac{d}{dt}\lambda_1(t) &= -\lambda_0(t) && (70)\\ \lambda_1(t) &= \int_t^1 \lambda_0(s)ds && (71)\\ &= \lambda_0(1)\frac{1}{\|\xi\|}\sin\big((1-t)\|\xi\|\big) && (72)\\ &\quad - \frac{1}{\|\xi\|^2}\big(*\xi \times \lambda_0(1)\big)\big(1 - \cos\big((1-t)\|\xi\|\big)\big) \\ &\quad + \frac{1}{\|\xi\|^3} * \xi\big(*\xi \cdot \lambda_0(1)\big)\big(1 - t - \sin\big((1-t)\|\xi\|\big)\big). \end{aligned}$$

6.3 Polynomial Regression

We apply a method similar to that of the previous section to derive an adjoint optimization scheme for Riemannian polynomial regression in a Lie group with right invariant metric. A variation of the first equation gives Eq. (46). Taking variations of the other equations, noting that $\overline{\nabla}$ is linear in each argument, we have

$$\frac{d}{dt}\delta\xi_i = \overline{\nabla}_{\delta\xi_1}\xi_i + \overline{\nabla}_{\xi_1}\delta\xi_i + \delta\xi_{i+1}. \tag{73}$$

Along with Eq. (46), these provide the essential equations for a polynomial perturbation Z of γ, which can be considered a kind of higher-order Jacobi field. Introducing adjoint variables $\lambda_0, \ldots, \lambda_k \in \mathfrak{g}$, the adjoint system is (see Appendix C for derivation)

$$\frac{d}{dt}\lambda_0 = -\operatorname{ad}^\dagger_{\xi_1}\lambda_0 \tag{74}$$

$$\frac{d}{dt}\lambda_1 = -\lambda_0 - \mathrm{sym}^\dagger_{\xi_1}\lambda_1 + \sum_{i=2}^{k}\left(-\overline{\nabla}_{\xi_i} - \mathrm{sym}^\dagger_{\xi_i}\right)\lambda_i \tag{75}$$

$$\frac{d}{dt}\lambda_i = -\lambda_{i-1} + \overline{\nabla}_{\xi_1}\lambda_i, \quad i = 2,\ldots,k, \tag{76}$$

or, using only ad and $\mathrm{ad}^\dagger$, as

$$\frac{d}{dt}\lambda_0 = -\mathrm{ad}^\dagger_{\xi_1}\lambda_0 \tag{77}$$

$$\frac{d}{dt}\lambda_1 = -\lambda_0 + \mathrm{ad}_{\xi_1}\lambda_1 - \mathrm{ad}^\dagger_{\lambda_1}\xi_1 + \frac{1}{2}\sum_{i=2}^{k}\left(\mathrm{ad}_{\xi_i}\lambda_i + \mathrm{ad}^\dagger_{\xi_i}\lambda_i - \mathrm{ad}^\dagger_{\lambda_i}\xi_i\right) \tag{78}$$

$$\frac{d}{dt}\lambda_i = -\lambda_{i-1} + \frac{1}{2}\left(\mathrm{ad}_{\xi_1}\lambda_i - \mathrm{ad}^\dagger_{\xi_1}\lambda_i - \mathrm{ad}^\dagger_{\lambda_i}\xi_1\right). \tag{79}$$

For $i = 2,\ldots,k$, these equations resemble the original polynomial equations. However, the evolution of λ_1 is influenced by all adjoint variables and higher-order velocities in a non-trivial way. The first adjoint equation again resembles the Euler-Poincaré equation, and its solution is given by Eq. (54).

6.3.1 *Polynomial Regression in* SO(3)

Revisiting the rotation group, we can extend the geodesic regression results to polynomials. Representing Lie algebra elements as 3-vectors ξ_i, the equations for higher order polynomials in SO(3) are

$$\frac{d}{dt}\gamma(t) = \left(*\xi_1(t)\right)\gamma(t) \tag{80}$$

$$\frac{d}{dt}\xi_1(t) = \xi_2(t) \tag{81}$$

$$\frac{d}{dt}\xi_i(t) = \frac{1}{2}\xi_1(t)\times\xi_i(t) + \xi_{i+1}(t), \quad i = 2,\ldots,k-1 \tag{82}$$

$$\frac{d}{dt}\xi_k(t) = \frac{1}{2}\xi_1(t)\times\xi_k(t). \tag{83}$$

In this case, closed form integration isn't available, even with a biinvariant metric. Even for higher order polynomials, the first adjoint equation is integrated in closed form, giving

$$\lambda_0(t) = \gamma(t)^T\gamma(1)\lambda_0(1). \tag{84}$$

7 Lie Group Actions

So far, we've seen that polynomial regression is particularly convenient in Lie groups with right invariant metrics, reducing the adjoint system from second to first order using the closed form integral of λ_0. We now consider the case when a Lie group G acts on another manifold M which is itself equipped with a Riemannian metric. For our purposes, the group action need not be transitive, in which case the target space is called a "homogeneous space" for G.

Although the two approaches sometimes coincide, generally one must choose between using polynomials defined by the metric in M, ignoring the action of G, or using curves defined by the action of polynomials in G on points in M. In cases when a Riemannian Lie group is known to act on the space M, the primary object of interest is usually not the path in the object space M, but the path of symmetries described by the group elements. Therefore it is most natural to use the Lie group structure to define paths in object space. We employ this approach, in which polynomial regression under a Riemannian Lie group action is studied primarily using the Lie group elements.

Following this plan, we model a polynomial in M as a curve $p(t)$ defined using the group action:

$$p(t) = \gamma(t).p_0 \tag{85}$$

where γ is a polynomial of order k in G with parameters

$$\gamma(0) \in G, \qquad \xi_1,\ldots,\xi_k \in \mathfrak{g} \tag{86}$$

and $p_0 \in M$ is a base point in the object space. Invariance of the metric on G allows us to assume, without loss of flexibility in the model, that the base deformation is the identity: $\gamma(0) = e \in G$. Optimization is done by fixing $\gamma(0) = e \in G$ and minimizing a least squares objective function defined using the metric on M, with respect to the base point $p_0 \in M$ and the parameters of the Lie group polynomial, $\xi_1,\ldots,\xi_k \in \mathfrak{g}$. This is accomplished using a similar adjoint method to that presented in the previous sections, but where the jump discontinuities in λ_0 are modified due to this change in objective function. In the following sections, we discuss this in more detail and also derive the gradients with respect to the base point p_0.

7.1 Action on a General Manifold

A smooth group action can be differentiated to obtain a mapping from the Lie algebra $\mathfrak{g}$ to the tangent space T_pM at any point $p \in M$. Given a curve $g(t) : (-\epsilon,\epsilon) \to G$ such that $g(0) = e$ and $\frac{d}{dt}|_{t=0}g(t) = \xi \in \mathfrak{g}$, define the following mapping (c.f. [16]):

$$\rho_p(\xi) := \left.\frac{d}{dt}\right|_{t=0} g(t).p. \tag{87}$$

The function ρ_p is a linear mapping from $\mathfrak{g}$ to T_pM, and as such it has a dual $\rho_p^* : T_p^*M \to \mathfrak{g}^*$ that maps cotangent vectors in M to the Lie coalgebra $\mathfrak{g}^*$. This dual mapping

we refer to as the cotangent lift momentum map and use the notation $\mathbf{J} : T^*M \to \mathfrak{g}^*$.

The most important property of $\mathbf{J}$ is that it is preserved under the coadjoint action:

$$\mathrm{Ad}^*_g \mathbf{J}m = \mathbf{J}g.m \quad \forall m \in T^*M. \tag{88}$$

The action of g on the cotangent bundle, which appears on the right-hand side above, maps a cotangent vector μ at point p to the vector $g.\mu \in T^*_{g.p}M$. Replacing squared norm with squared geodesic distance on the Riemannian manifold M, the first adjoint variable is then given by

$$\lambda_0(t) = \sum_{j,t_j>t} \mathbf{J}\gamma(t)\gamma(t_j)^{-1}.(\mathrm{Log}_{\gamma(t_j).p_0} J_j)^\flat. \tag{89}$$

Of particular interest is the case when the metric on G and the metric on the manifold M coincide, in the sense that for any vectors $\xi, \mu \in \mathfrak{g}$ and points $p \in M$:

$$\langle \xi, \mu \rangle_{\mathfrak{g}} = \langle \xi.p, \mu.p \rangle_{T_pM}. \tag{90}$$

Fixing a base point $p_0 \in M$, this means the mapping $g \to g.p_0$ is a Riemannian submersion. If, additionally, the metric on G is biinvariant, this implies that the covariant derivative satisfies [33]

$$\nabla_{\xi.p}\mu.p = (\overline{\nabla}_\xi \mu).p \tag{91}$$

so that geodesics and polynomials in M are generated by polynomials in G along with the action on the base point p_0.

7.1.1 Example: Rotations of the Sphere

Consider the sphere of radius one in $\mathbb{R}^3$, which is denoted $\mathbb{S}^2$. The group SO(3) acts naturally on the sphere. For this example, we will use the biinvariant metric on SO(3), which corresponds to using the identity for the A matrix in Sect. 6.2. Representing points on the sphere as unit vectors in $\mathbb{R}^3$, the group action is simply left multiplication by a matrix in SO(3):

$$\gamma.p = \gamma p \tag{92}$$

$$\xi.p = \xi p \tag{93}$$

for all $\gamma \in \mathrm{SO}(3), \xi \in \mathfrak{so}(3), p \in \mathbb{S}^2, v \in T_p\mathbb{S}^2$. The infinitesimal action is in fact a cross product, which is easily seen using the star map:

$$\xi.p = \xi p = (*\xi) \times p. \tag{94}$$

Representing elements in $\mathfrak{so}(3)^*$ as 3-vectors, we derive the cotangent lift momentum map as well; letting $a \in T_p\mathbb{S}^2$,

$$\mathbf{J}a = *(p \times a). \tag{95}$$

This can be interpreted as converting a linear momentum on the surface of the sphere into an angular momentum in $\mathfrak{so}(3)$ using the cross product with the moment arm p. The standard metric on the sphere corresponds to the standard biinvariant metric on SO(3) so that, as discussed previously, polynomials on $\mathbb{S}^2$ correspond to polynomials in SO(3) acting on points on the sphere.

The polynomial equations for the sphere are precisely those for SO(3), along with the action of $\gamma(t)$ on the base point $p_0 \in \mathbb{S}^2$. The derivative of $\gamma(t)$ is replaced by the equation

$$\frac{d}{dt}p(t) = \frac{d}{dt}\big(\gamma(t).p_0\big) = \xi_1(t).p(t). \tag{96}$$

The evolution of ξ_i is the same as that for SO(3). Figure 6 shows example polynomial curves in the rotation group and their action on a point on the sphere. Notice that the example polynomials on the sphere are precisely those shown in Fig. 1, although they were generated here using polynomials on SO(3) instead of integrating directly on the sphere.

In order to integrate the adjoint equations, the jump discontinuities must be computed using the log map on the sphere:

$$\mathrm{Log}_x\, y = \theta\left(\frac{y - \cos\theta x}{\sin\theta}\right), \quad \cos\theta = x^T y. \tag{97}$$

The flatting operation acts trivially on this vector, and the action of SO(3) on covectors corresponds to matrix-vector multiplication. Using this, along with the momentum map $\mathbf{J}$, we have the jump discontinuities for the first adjoint variable λ_0:

$$\lambda_0\big(t_j^-\big) - \lambda_0\big(t_j^+\big) = \gamma(t_j) \times (\mathrm{Log}_{\gamma(t_j)} J_j). \tag{98}$$

The higher adjoint variables satisfy the same ODEs as in Sect. 6.3.1.

7.2 Lie Group Actions on Vector Spaces

We will assume in this section that the manifold is a vector space V and that G acts linearly on the left on V. Given a smooth linear group action, a vector ξ in the Lie algebra $\mathfrak{g}$ acts linearly on a vector $v \in V$ in the following way

$$\xi.v = \frac{d}{d\epsilon}\bigg|_{\epsilon=0} g(\epsilon).v \tag{99}$$

where $g(\epsilon)$ is a curve in G satisfying $g(0) = e$ and $\frac{d}{d\epsilon}|_{\epsilon=0} g(\epsilon) = \xi$. Again we use the notation $\rho_v : \mathfrak{g} \to V$ to denote right-multiplication under this action:

$$\rho_v \xi := \xi.v \quad \forall v \in V, \xi \in \mathfrak{g}. \tag{100}$$

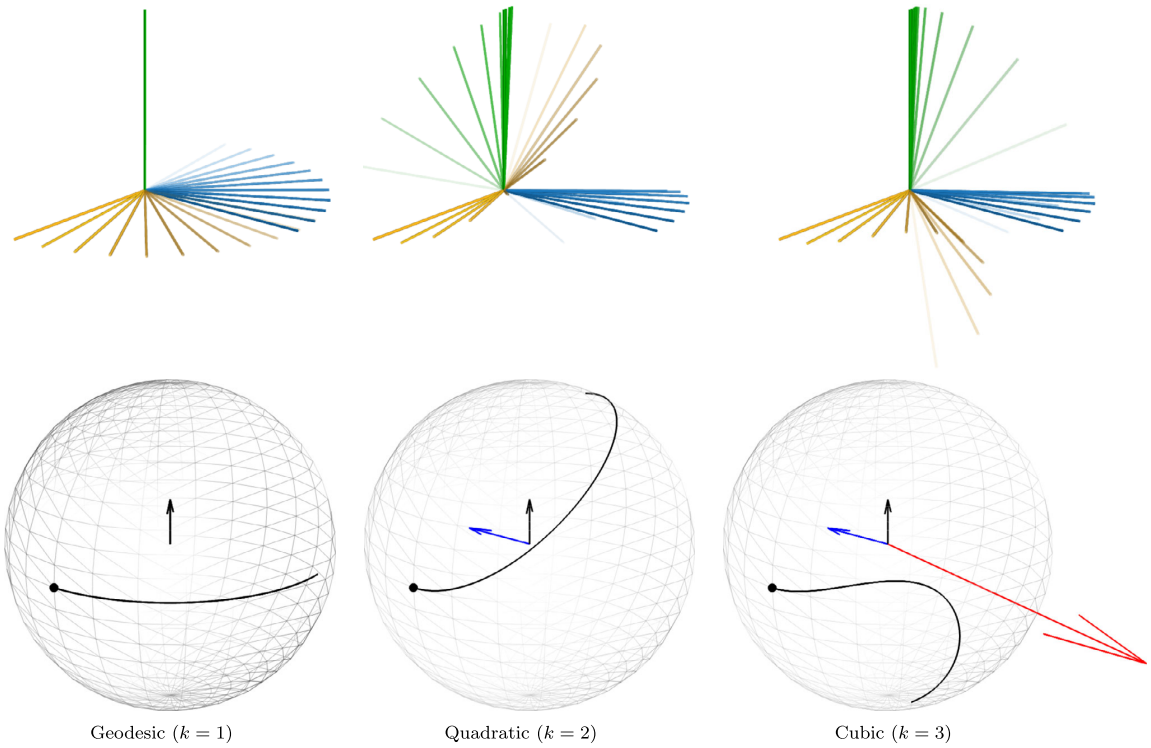


Fig. 6 Sample polynomial curves in SO(3) and their action on a base point $p_0 \in \mathbb{S}^2$ (*black dot*) on the sphere. In the *top row*, the rotating coordinate axes are shown for three polynomials. In the *bottom row*, the *arrows* show the vectors $\xi_1(0)$ (*black*), $\xi_2(0)$ (*blue*), and $\xi_3(0)$ (*red*), representing initial angular velocity, acceleration, and jerk. The action on the base point, $p(t) = \gamma(t).p_0 \in \mathbb{S}^2$, is represented as a *black curve* on the sphere. A geodesic corresponds to constant angular velocity, while the non-zero acceleration and jerk in the quadratic and cubic curves tilt the rotation axis

In the vector space setting, the cotangent lift momentum map (again defined as the dual of ρ_v), is written using the diamond notation introduced in [16]:

$$v \diamond a \in \mathfrak{g}^* \quad \forall v \in V, a \in V^*, \tag{101}$$

$$(v \diamond a, \xi)_{(\mathfrak{g}*,\mathfrak{g})} := (a, \rho_v \xi)_{(V^*,V)} \quad \forall \xi \in \mathfrak{g}. \tag{102}$$

The diamond map interacts with the coadjoint action Ad^* in a convenient way:

$$\mathrm{Ad}^*_{g^{-1}}(v \diamond a) = (g.v) \diamond (g.a). \tag{103}$$

This relation is fundamental in that it shows that the diamond map is preserved under the coadjoint action. This is quite useful in our case, as we will soon see that diamond maps show up commonly in variational problems on inner product spaces.

Commonly, data is provided in the form of points J_i in the vector space V. In that case, the inner product on V is used to write the regression problem as a minimization of

$$E(\gamma, v_0) = \frac{1}{2} \sum_{j=1}^{N} \left\| \gamma(t_j).v_0 - J_j \right\|_V^2, \tag{104}$$

subject to the constraint that γ is a polynomial in G and $v_0 \in V$ is an evolving template vector. Without loss of generality, $\gamma(0)$ can also be constrained to be the identity so that v_0 is the template vector at time zero. Optimization of Eq. (104) with respect to v_0 requires the variation

$$\delta_{v_0} E = \sum_{j=1}^{N} \gamma(t_j)^{-1} \big(\gamma(t_j).v_0 - J_j\big)^\flat. \tag{105}$$

Here the musical flat symbol $\flat$ denotes lowering of indices using the metric on V, an operation mapping V to V^*. If the group G acts by isometries on V, then the group action commutes with flatting and the optimal base vector v_0 can

be computed in closed form

$$\hat{v}_0 = \frac{1}{N}\sum_{j=1}^{N} \gamma(t_j)^{-1}.J_j. \tag{106}$$

Even when G does not act by isometries, the optimal base vector can often be solved for in closed form.

The variation with respect to $\gamma(t_j)$ is more interesting:

$$\delta_{\gamma(t_j)} E = \big(\gamma(t_j).v_0\big) \diamond \big(\gamma(t_j).v_0 - J_j\big)^\flat. \tag{107}$$

Using this along with the relation between the coadjoint action and diamond map, we can write the first polynomial adjoint variable in closed form

$$\lambda_0(t) = \sum_{j,t_j>t} \big(\gamma(t).v_0\big) \diamond \big(\gamma(t)\gamma(t_j)^{-1}.\big(\gamma(t_j).v_0 - J_j\big)^\flat\big). \tag{108}$$

7.2.1 Example: Diffeomorphically Deforming Images

Right invariant Sobolev metrics on groups of diffeomorphisms are the main objects of study in computational anatomy [30]. Describing an image I as a square integrable function of a domain $\Omega \subset \mathbb{R}^d$, the left action of a diffeomorphism $\gamma \in \mathrm{Diff}(\Omega)$ is

$$\gamma.I = I \circ \gamma^{-1}. \tag{109}$$

The corresponding infinitesimal action of a velocity field ξ on an image is

$$\xi.I = -\xi^T \nabla I \tag{110}$$

and the diamond map is

$$(I \diamond \alpha)(y) = -\alpha(y)\nabla I(y). \tag{111}$$

Geodesic regression in this context, using an adjoint optimization method, has been previously studied [31]. Using their method, the initial momentum of a geodesic is constrained by horizontal: that is, $L\xi_1(0) = I_0 \diamond \alpha(0)$. As a result, changes in base image I_0 influence the behavior of the deformation itself.

Using our method, the base velocity vectors ξ_i are not constrained to be horizontal. Implementation of polynomial regression involves the expression above for the diamond map, along with the ad and ad* operators [28]

$$\mathrm{ad}_\xi X = \mathrm{D}\xi X - \mathrm{D}X\xi, \tag{112}$$

$$\mathrm{ad}^*_\xi m = \mathrm{D}m\xi + m\,\mathrm{div}\,\xi + (\mathrm{D}\xi)^T m. \tag{113}$$

Inserting this into the right Euler-Poincaré equation yields the well-known EPDiff equation for geodesic evolution in the diffeomorphism group [16]:

$$\frac{d}{dt}m = -\mathrm{D}m\xi - m\,\mathrm{div}\,\xi - (\mathrm{D}\xi)^T m. \tag{114}$$

For polynomials, momenta $m_i = L\xi_i$ are introduced and this EPDiff equation is generalized to

$$\frac{d}{dt}m_1 = -\mathrm{D}m_1\xi_1 - m_1\,\mathrm{div}\,\xi_1 - (\mathrm{D}\xi_1)^T m_1 + m_2 \tag{115}$$

$$\begin{aligned}\frac{d}{dt}m_i = m_{i+1} + \frac{1}{2}\big(&L(\mathrm{D}\xi_1\xi_i - \mathrm{D}\xi_i\xi_1)\\ &- \mathrm{D}m_i\xi_1 - (\mathrm{D}\xi_1)^T m_i - m_i\,\mathrm{div}\,\xi_1\\ &- \mathrm{D}m_1\xi_i - (\mathrm{D}\xi_i)^T m_1 - m_1\,\mathrm{div}\,\xi_i\big)\end{aligned} \tag{116}$$

$$\begin{aligned}\frac{d}{dt}m_k = \frac{1}{2}\big(&L(\mathrm{D}\xi_1\xi_i - \mathrm{D}\xi_i\xi_1)\\ &- \mathrm{D}m_k\xi_1 - (\mathrm{D}\xi_1)^T m_k - m_k\,\mathrm{div}\,\xi_1\\ &- \mathrm{D}m_1\xi_k - (\mathrm{D}\xi_i)^T m_1 - m_1\,\mathrm{div}\,\xi_k\big)\end{aligned} \tag{117}$$

The estimation of the base image I_0 is simplified, as Eq. (105) is solved in closed form using

$$I_0(y) = \frac{\sum_j |\mathrm{D}\gamma_j(y)| J_j \circ \gamma_j(y)}{\sum_j |\mathrm{D}\gamma_j(y)|}. \tag{118}$$

As an example of image regression, synthetic data were generated and geodesic regression was performed using the adjoint method described above. Figure 7 shows the input images, as well as the estimated geodesic trend, which matches the input data well. Note that although the method presented in [31] is similar, using our abstraction, geodesic regression can be generalized to polynomials of any order, and to data which are not necessarily scalar-valued images.

8 Discussion

The Riemannian polynomial framework we have presented provides a general approach to regression for manifold-valued data. The greatest limitation to performing polynomial regression on a general Riemannian manifold is that it requires computation of the Riemannian curvature tensor, which is often tedious [29]. In a Lie group or homogeneous space, we have shown that the symmetries provided by the group allow for not only simple integration using parallel transport in the Lie algebra, but also simplified adjoint equations that do not require explicit curvature computation.

The theory of rolling maps on the sphere, introduced by Jupp & Kent [18], offer another perspective on Riemannian polynomials. On the sphere, this interesting interpretation is

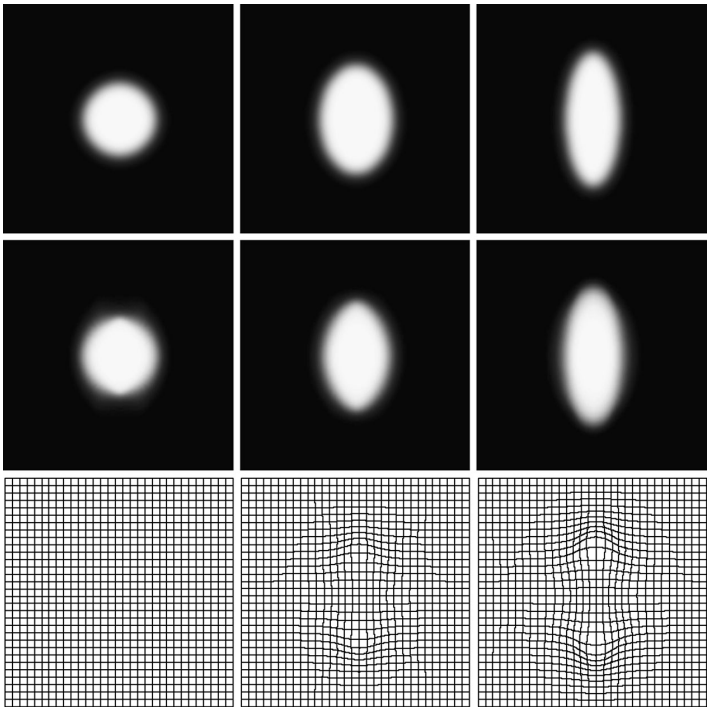


Fig. 7 Image regression example. Three synthetic images where generated (*top row*) at times 0,0.5,1. Geodesic regression was performed, resulting in the images shown in the *second row*, corresponding to the deformations in the *last row*

related to the group action described above. Given a curve $\gamma : [0, 1] \to \mathbb{S}^2$, consider embedding both the sphere and a plane in $\mathbb{R}^3$ such that the plane is tangent to the sphere at the point $\gamma(0)$. Now roll the sphere along so that it remains tangent at $\gamma(t)$ at every time, and such that no slipping or twisting occurs. The resulting path, $\gamma_u : [0, 1] \to \mathbb{R}^2$, traced out on the plane is called the unwrapped curve. Remarkably, the property that γ is a k-order polynomial on $\mathbb{S}^2$ is equivalent to the unwrapped curve γ_u being a k-order polynomial in the conventional sense. For more information regarding this connection to Jupp & Kent's rolling maps, as well as a comparison to Noakes' cubic splines [32], the reader is referred to the literature of Leite & Krakowski [24].

Appendix A: Numerical Integration of the Polynomial Equations

By definition, in the limit $\Delta t \to 0$, the exponential map satisfies $\dot{\gamma}(t) = v_1(t)$. To see that the forward integration algorithm shown in Algorithm 1 approximates the polynomial equations, let $w(t)$ be any vector field *parallel* along $\gamma(t)$. That is,

$$\nabla_{\dot{\gamma}(t)} w(t) = 0. \tag{119}$$

Denote by $P_{\Delta t}(t) = \mathrm{ParTrans}(p, \Delta t v, w)$ the parallel transport of a vector $w \in T_p M$ along a geodesic from point p for time Δt in the direction of vector $v \in T_p M$. Then

$$\frac{d}{dt}\langle w, v_i\rangle = \langle \nabla_{\dot{\gamma}} w, v_i\rangle + \langle w, \nabla_{\dot{\gamma}} v_i\rangle = \langle w, \nabla_{\dot{\gamma}} v_i\rangle \tag{120}$$

Now consider approximation of this inner product derivative under our integration scheme:

$$\begin{aligned}\frac{d}{dt}\langle w, v_i\rangle \approx \lim_{\Delta t\to 0} \frac{1}{\Delta t}\big(&\big\langle P_{\Delta t} w(t),\, P_{\Delta t}\big(v_i(t) + \Delta t v_{i+1}(t)\big)\big\rangle \\ &- \big\langle w(t), v_i(t)\big\rangle\big).\end{aligned} \tag{121}$$

The parallel transport operator is linear in the vectors being transported, so

$$\begin{aligned}\frac{d}{dt}\langle w, v_i\rangle &\approx \lim_{\Delta t\to 0} \frac{1}{\Delta t}\big(\big\langle P_{\Delta t} w(t),\, P_{\Delta t} v_i(t)\big\rangle \\ &\quad + \Delta t\big\langle P_{\Delta t} w(t), v_{i+1}(t)\big\rangle - \big\langle w(t), v_i(t)\big\rangle\big) \\ &= \lim_{\Delta t\to 0} \frac{1}{\Delta t}\big(\big(\big\langle P_{\Delta t} w(t),\, P_{\Delta t} v_i(t)\big\rangle - \big\langle w(t), v_i(t)\big\rangle\big) \\ &\quad + \lim_{\Delta t\to 0}\big\langle P_{\Delta t} w(t), v_{i+1}(t)\big\rangle\big)\end{aligned} \tag{122}$$

The first line is zero, by definition of parallel transport. Also note that $\lim_{\Delta t\to 0} P_{\Delta t} w = w$, so that

$$\frac{d}{dt}\langle w, v_i\rangle = \langle w, \nabla_{\dot\gamma} v_i\rangle \approx \langle w, v_{i+1}\rangle. \tag{123}$$

As this holds for any parallel vector field w, this implies that our integration algorithm approximates the polynomial equation

$$\nabla_{\dot\gamma} v_i = v_{i+1}. \tag{124}$$

Appendix B: Derivation of Adjoint Equations in Riemannian Manifolds

In this appendix we derive the adjoint system for the polynomial regression problem. The approach to calculus of variations on Riemannian manifolds described here is very similar to that employed by Noakes et al. [32]. Consider a simplified objective function containing only a single data term, at time T:

$$\begin{aligned} E\big(\gamma, \{v_i\}, \{\lambda_i\}\big) = d\big(\gamma(T), y\big)^2 &+ \int_0^T \langle \lambda_0, \dot\gamma - v_1\rangle dt \\ &+ \sum_{i=1}^{k-1} \int_0^T \langle \lambda_i, \nabla_{\dot\gamma} v_i - v_{i+1}\rangle dt \\ &+ \int_0^T \langle \lambda_k, \nabla_{\dot\gamma} v_k\rangle dt. \end{aligned} \tag{125}$$

Now consider taking variations of E with respect to the vector fields v_i. For each i there are only two terms containing v_i, so if W is a test vector field along γ, then the variation of E with respect to v_i in the direction W satisfies

$$\int_0^T \langle \delta_{v_i} E, W\rangle dt = \int_0^T \langle \lambda_i, \nabla_{\dot\gamma} W\rangle dt - \int_0^T \langle \lambda_{i-1}, W\rangle dt. \tag{126}$$

The first term is integrated by parts to yield

$$\begin{aligned} \int_0^T \langle \delta_{v_i} E, W\rangle dt = \langle \lambda_i, W\rangle|_0^T &- \int_0^T \langle \nabla_{\dot\gamma}\lambda_i, W\rangle dt \\ &- \int_0^T \langle \lambda_{i-1}, W\rangle dt. \end{aligned} \tag{127}$$

The variation with respect to v_i for $i = 1, \ldots, k$ is then given by

$$\delta_{v_i(t)} E = 0 = -\nabla_{\dot\gamma}\lambda_i - \lambda_{i-1}, \quad t \in (0, T) \tag{128}$$

$$\delta_{v_i(T)} E = 0 = \lambda_i(T) \tag{129}$$

$$\delta_{v_i(0)} E = -\lambda_i(t). \tag{130}$$

In order to determine the differential equation for λ_0, the variation with respect to γ must be computed. Let W again denote a test vector field along γ. For some $\epsilon > 0$, let $\{\gamma_s : s \in (-\epsilon, \epsilon)\}$ be a differentiable family of curves satisfying

$$\gamma_0 = \gamma \tag{131}$$

$$\frac{d}{ds}\gamma_s\Big|_{s=0} = W. \tag{132}$$

If ϵ is chosen small enough, the vector field W can be extended to a neighborhood of γ such that $[W, \dot\gamma_s] = 0$, where a dot indicates the derivative in the $\frac{\partial}{\partial t}$ direction. The vanishing Lie bracket implies the following identities

$$\nabla_W \dot\gamma_s = \nabla_{\dot\gamma_s} W \tag{133}$$

$$\nabla_W \nabla_{\dot\gamma_s} = \nabla_{\dot\gamma_s}\nabla_W + R(W, \dot\gamma_s). \tag{134}$$

Finally, the vector fields v_i, λ_i are extended along γ_s via parallel translation, so that

$$\nabla_W v_i = 0 \tag{135}$$

$$\nabla_W \lambda_i = 0. \tag{136}$$

The variation of E with respect to γ satisfies

$$\begin{aligned} \int_0^T \langle \delta_\gamma E, W\rangle dt &= \frac{d}{ds} E\big(\gamma_s, \{v_i\}, \{\lambda_i\}\big)\big|_{s=0} \\ &= -\big\langle \mathrm{Log}_{\gamma(T)} y, W(T)\big\rangle \\ &\quad + \frac{d}{ds}\int_0^T \langle \lambda_0, \dot\gamma_s - v_1\rangle dt\Big|_{s=0} \\ &\quad + \frac{d}{ds}\sum_{i=1}^{k-1}\int_0^T \langle \lambda_i, \nabla_{\dot\gamma_s} v_i - v_{i+1}\rangle dt\Big|_{s=0} \\ &\quad + \frac{d}{ds}\int_0^T \langle \lambda_k, \nabla_{\dot\gamma_s} v_k\rangle dt\Big|_{s=0}. \end{aligned} \tag{137}$$

As the λ_i are extended via parallel translation, their inner products satisfy

$$\frac{d}{ds}\langle \lambda_i, U\rangle|_{s=0} = \langle \nabla_W \lambda_i, U\rangle + \langle \lambda_i, \nabla_W U\rangle = \langle \lambda_i, \nabla_W U\rangle. \tag{138}$$

Then applying this to each term in the previous equation,

$$\begin{aligned} \int_0^T \langle \delta_\gamma E, W\rangle dt = &-\big\langle \mathrm{Log}_{\gamma(T)} y, W(T)\big\rangle \\ &+ \int_0^T \langle \lambda_0, \nabla_W\dot\gamma - \nabla_W v_1\rangle dt \\ &+ \sum_{i=1}^{k-1}\int_0^T \langle \lambda_i, \nabla_W\nabla_{\dot\gamma} v_i - \nabla_W v_{i+1}\rangle dt \\ &+ \int_0^T \langle \lambda_k, \nabla_W\nabla_{\dot\gamma} v_k\rangle dt. \end{aligned} \tag{139}$$

Then by construction, since $\nabla_W v_i = 0$,

$$\int_0^T \langle \delta_\gamma E, W\rangle dt = -\big\langle \mathrm{Log}_{\gamma(T)} y, W(T)\big\rangle + \int_0^T \langle \lambda_0, \nabla_W \dot\gamma\rangle dt + \sum_{i=1}^k \int_0^T \langle \lambda_i, \nabla_W \nabla_{\dot\gamma} v_i\rangle dt. \quad (140)$$

Then using the Lie bracket and curvature identities, this is written as

$$\int_0^T \langle \delta_\gamma E, W\rangle dt = -\big\langle \mathrm{Log}_{\gamma(T)} y, W(T)\big\rangle + \int_0^T \langle \lambda_0, \nabla_{\dot\gamma} W\rangle dt + \sum_{i=1}^k \int_0^T \big\langle \lambda_i, \nabla_{\dot\gamma}\nabla_W v_i + R(W,\dot\gamma)v_i\big\rangle dt, \quad (141)$$

which is further simplified, again using the identity $\nabla_W v_i = 0$:

$$\int_0^T \langle \delta_\gamma E, W\rangle dt = -\big\langle \mathrm{Log}_{\gamma(T)} y, W(T)\big\rangle + \int_0^T \langle \lambda_0, \nabla_{\dot\gamma} W\rangle dt + \sum_{i=1}^k \int_0^T \big\langle \lambda_i, R(W,\dot\gamma)v_i\big\rangle dt, \quad (142)$$

Using the Bianchi identities, it can be demonstrated that the curvature tensor satisfies the identity [9]:

$$\big\langle A, R(B,C)D\big\rangle = -\big\langle B, R(D,A)C\big\rangle, \quad (143)$$

for any vectors A, B, C, D. The covariant derivative along γ is also integrated by parts to arrive at

$$\int_0^T \langle \delta_\gamma E, W\rangle dt = -\big\langle \mathrm{Log}_{\gamma(T)} y, W(T)\big\rangle + \langle \lambda_0, W\rangle|_0^T - \int_0^T \langle \nabla_{\dot\gamma}\lambda_0, W\rangle dt - \sum_{i=1}^k \int_0^T \big\langle R(v_i,\lambda_i)\dot\gamma, W\big\rangle dt. \quad (144)$$

Finally, gathering terms, the adjoint equation for λ_0 and its gradients are obtained:

$$\delta_{\gamma(t)} E = 0 = -\nabla_{\dot\gamma}\lambda_0 - \sum_{i=1}^k R(v_i,\lambda_i)\dot\gamma, \quad t \in (0,T) \quad (145)$$

$$\delta_{\gamma(T)} E = 0 = -\mathrm{Log}_{\gamma(T)} y + \lambda_0 \quad (146)$$

$$\delta_{\gamma(0)} E = -\lambda_0. \quad (147)$$

Along with the variations with respect to v_i, this constitutes the full adjoint system. Extension to the case of multiple data at multiple time points is trivial, and results in the adjoint system presented in Sect. 4.

Appendix C: Derivation of Adjoint Equations in Lie Groups

Let G be a Lie group with Lie algebra $\mathfrak{g}$, equipped with a right invariant metric. Let $\gamma : [0,1] \to G$ be a polynomial in G of order k with right-trivialized velocities $\xi_i : [0,1] \to \mathfrak{g}$. Recall the equations for a perturbation $Z, \delta\xi_i$ of this polynomial:

$$\frac{d}{dt} Z = \delta\xi_1 - \mathrm{ad}_{\xi_1} Z \quad (148)$$

$$\frac{d}{dt}\delta\xi_i = \overline{\nabla}_{\delta\xi_1}\xi_i + \overline{\nabla}_{\xi_1}\delta\xi_i + \delta\xi_{i+1}. \quad (149)$$

The second equation can be rewritten

$$\frac{d}{dt}\delta\xi_i = \frac{1}{2}\mathrm{ad}_{\delta\xi_1}\xi_i + \frac{1}{2}\mathrm{sym}_{\delta\xi_1}\xi_i + \overline{\nabla}_{\xi_1}\delta\xi_i + \delta\xi_{i+1} \quad (150)$$

$$= -\frac{1}{2}\mathrm{ad}_{\xi_i}\delta\xi_1 + \frac{1}{2}\mathrm{sym}_{\xi_i}\delta\xi_1 + \overline{\nabla}_{\xi_1}\delta\xi_i + \delta\xi_{i+1} \quad (151)$$

$$= (-\overline{\nabla}_{\xi_i} + \mathrm{sym}_{\xi_i})\delta\xi_1 + \overline{\nabla}_{\xi_1}\delta\xi_i + \delta\xi_{i+1}. \quad (152)$$

This suggests the following matrix form ODE:

$$\frac{d}{dt}\begin{pmatrix} Z \\ \delta\xi_1 \\ \vdots \\ \delta\xi_k \end{pmatrix} = \begin{pmatrix} \mathrm{ad}_{\xi_1} & I & \cdots & 0 & 0 & 0 \\ 0 & \mathrm{sym}_{\xi_1} & I & \cdots & 0 & \\ 0 & -\overline{\nabla}_{\xi_2} + \mathrm{sym}_{\xi_2} & \overline{\nabla}_{\xi_1} & I & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & -\overline{\nabla}_{\xi_k} + \mathrm{sym}_{\xi_k} & 0 & \cdots & & \overline{\nabla}_{\xi_1} \end{pmatrix} \times \begin{pmatrix} Z \\ \delta\xi_1 \\ \vdots \\ \delta\xi_k \end{pmatrix}. \quad (153)$$

In order to derive the adjoint Jacobi field, one simply computes the negative adjoint of the matrix in the above equa-

tion. The adjoint of the above matrix is

$$\begin{pmatrix} -\operatorname{ad}^\dagger_{\xi_1} & 0 & \cdots & 0 & 0 \\ -I & -\operatorname{sym}^\dagger_{\xi_1} & \overline{\nabla}^\dagger_{\xi_2} - \operatorname{sym}^\dagger_{\xi_2} & \cdots & \overline{\nabla}^\dagger_{\xi_k} - \operatorname{sym}^\dagger_{\xi_k} \\ 0 & -I & -\overline{\nabla}_{\xi_1} & 0 & \cdots \\ \vdots & & & & 0 \\ 0 & 0 & \cdots & -I & -\overline{\nabla}^\dagger_{\xi_1} \end{pmatrix}. \tag{154}$$

Now note that the adjoint of the $\overline{\nabla}_\xi$ operator is $-\overline{\nabla}_\xi$, since (using Eq. (52))

$$2\overline{\nabla}^\dagger_X Y = \operatorname{ad}^\dagger_X Y + \operatorname{sym}^\dagger_X Y \tag{155}$$

$$= \operatorname{ad}^\dagger_X Y - \operatorname{ad}_X Y + \operatorname{ad}^\dagger_Y X \tag{156}$$

$$= -\operatorname{ad}_X Y - \operatorname{sym}_X Y \tag{157}$$

$$= -2\overline{\nabla}_X Y. \tag{158}$$

Now let $\lambda_0, \ldots, \lambda_k : [0,1] \to \mathfrak{g}$ be adjoint variables representing gradients with respect to position γ and velocities $\xi_1, \ldots, \xi_k$. Using the equations above, we write the reduced polynomial adjoint equations as

$$\frac{d}{dt}\lambda_0 = -\operatorname{ad}^\dagger_{\xi_1} \lambda_0 \tag{159}$$

$$\frac{d}{dt}\lambda_1 = -\lambda_0 - \operatorname{sym}^\dagger_{\xi_1} \lambda_1 + \sum_{i=2}^{k} \left(-\overline{\nabla}_{\xi_k} - \operatorname{sym}^\dagger_{\xi_k}\right)\lambda_k \tag{160}$$

$$\frac{d}{dt}\lambda_i = -\lambda_{i-1} + \overline{\nabla}_{\xi_1}\lambda_i \quad i = 2, \ldots, k. \tag{161}$$

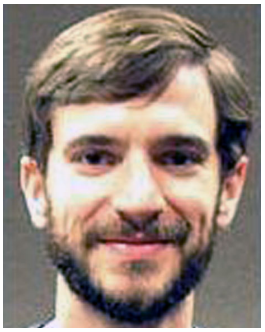
The first adjoint variable, λ_0, takes on jump discontinuities when passing data points, which are derived identically to the geodesic case. Also note that this derivation is for right invariant metrics using right trivialized vectors, but the equivalent derivation in the case of left invariance is essentially identical.

References

1. Bandulasiri, A., Gunathilaka, A., Patrangenaru, V., Ruymgaart, F., Thompson, H.: Nonparametric shape analysis methods in glaucoma detection. I. J. Stat. Sci. **9**, 135–149 (2009)
2. Bookstein, F.L.: Morphometric Tools for Landmark Data: Geometry and Biology. Cambridge Univ. Press, Cambridge (1991)
3. Bou-Rabee, N.: Hamilton-Pontryagin integrators on Lie groups. Ph.D. thesis, California Institute of Technology (2007)
4. Bruveris, M., Gay-Balmaz, F., Holm, D., Ratiu, T.: The momentum map representation of images. J. Nonlinear Sci. **21**(1), 115–150 (2011)
5. Burnham, K., Anderson, D.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York (2002)
6. Cates, J., Fletcher, P.T., Styner, M., Shenton, M., Whitaker, R.: Shape modeling and analysis with entropy-based particle systems. In: Proceedings of Information Processing in Medical Imaging (IPMI) (2007)
7. Cheeger, J., Ebin, D.G.: Comparison Theorems in Riemannian Geometry, vol. 365. AMS Bookstore, Providence (1975)
8. Davis, B.C., Fletcher, P.T., Bullitt, E., Joshi, S.C.: Population shape regression from random design data. Int. J. Comput. Vis. **90**(2), 255–266 (2010)
9. do Carmo, M.P.: Riemannian Geometry, 1st edn. Birkhäuser, Boston (1992)
10. Driesen, N, Raz, N: The influence of sex, age, and handedness on corpus callosum morphology: A meta-analysis. Psychobiology (1995). doi:10.3758/BF03332028
11. Dryden, I.L., Kume, A., Le, H., Wood, A.T.: A multi-dimensional scaling approach to shape analysis. Biometrika **95**(4), 779–798 (2008)
12. Fletcher, PT: Geodesic regression and the theory of least squares on Riemannian manifolds. Int. J. Comput. Vis. (2012). doi:10.1007/s11263-012-0591-y
13. Fletcher, P.T., Liu, C., Pizer, S.M., Joshi, S.C.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imaging **23**(8), 995–1005 (2004)
14. Giambò, R., Giannoni, F., Piccione, P.: An analytical theory for Riemannian cubic polynomials. IMA J. Math. Control Inf. **19**(4), 445–460 (2002)
15. Hinkle, J., Muralidharan, P., Fletcher, P.T., Joshi, S.C.: Polynomial regression on Riemannian manifolds. In: ECCV, Florence, Italy, vol. 3, pp. 1–14 (2012)
16. Holm, D.D., Marsden, J.E., Ratiu, T.S.: The Euler-Poincaré equations and semidirect products with applications to continuum theories. Adv. Math. **137**, 1–81 (1998)
17. Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: geodesic principal component analysis for Riemannian manifolds modulo Lie group actions. Discussion paper with rejoinder. Stat. Sin. **20**, 1–100 (2010)
18. Jupp, P.E., Kent, J.T.: Fitting smooth paths to spherical data. Appl. Stat. **36**(1), 34–46 (1987)
19. Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Bull. Lond. Math. Soc. **16**(2), 81–121 (1984)
20. Kendall, D.G.: A survey of the statistical theory of shape. Stat. Sci. **4**(2), 87–99 (1989)
21. Kenobi, K., Dryden, I.L., Le, H.: Shape curves and geodesic modelling. Biometrika **97**(3), 567–584 (2010)
22. Kume, A., Dryden, I.L., Le, H.: Shape-space smoothing splines for planar landmark data. Biometrika **94**(3), 513–528 (2007). doi:10.1093/biomet/asm047
23. Le, H., Kendall, D.G.: The Riemannian structure of Euclidean shape spaces: a novel environment for statistics. Ann. Stat. **21**(3), 1225–1271 (1993)
24. Silva Leite, F, Krakowski, K: Covariant differentiation under rolling maps. Departamento de Matemática, Universidade of Coimbra, Portugal (2008), No. 08-22, 1–8
25. Lewis, A., Murray, R.: Configuration controllability of simple mechanical control systems. SIAM J. Control Optim. **35**(3), 766–790 (1997)
26. Machado, L., Leite, F.S., Krakowski, K.: Higher-order smoothing splines versus least squares problems on Riemannian manifolds. J. Dyn. Control Syst. **16**(1), 121–148 (2010)
27. Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)
28. Marsden, J., Ratiu, T.: Introduction to Mechanics and Symmetry: a Basic Exposition of Classical Mechanical Systems, vol. 17. Springer, Berlin (1999)

29. Micheli, M., Michor, P., Mumford, D.: Sectional curvature in terms of the cometric, with applications to the Riemannian manifolds of landmarks. SIAM J. Imaging Sci. **5**(1), 394–433 (2012)
30. Miller, M.I., Trouvé, A., Younes, L.: Geodesic shooting for computational anatomy. J. Math. Imaging Vis. **24**(2), 209–228 (2006). doi:10.1007/s10851-005-3624-0
31. Niethammer, M., Huang, Y., Vialard, F.X.: Geodesic regression for image time-series. In: Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) (2011)
32. Noakes, L., Heinzinger, G., Paden, B.: Cubic splines on curved surfaces. IMA J. Math. Control Inf. **6**, 465–473 (1989)
33. O'Neill, B.: The fundamental equations of a submersion. Mich. Math. J. **13**(4), 459–469 (1966)
34. Shi, X., Styner, M., Lieberman, J., Ibrahim, J.G., Lin, W., Zhu, H.: Intrinsic regression models for manifold-valued data. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2009, pp. 192–199. Springer, Berlin (2009)
35. Singh, N., Wang, A., Sankaranarayanan, P., Fletcher, P., Joshi, S.: Genetic, structural and functional imaging biomarkers for early detection of conversion from MCI to AD. In: Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 132–140. Springer, Berlin (2012)
36. Turaga, P., Veeraraghavan, A., Srivastava, A., Chellappa, R.: Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11), 2273–2286 (2011)
37. Vaillant, M., Glaunes, J.: Surface matching via currents. In: Information Processing in Medical Imaging, pp. 1–5. Springer, Berlin (2005)
38. Younes, L., Qiu, A., Winslow, R., Miller, M.: Transport of relational structures in groups of diffeomorphisms. J. Math. Imaging Vis. **32**(1), 41–56 (2008)
39. Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage **31**(3), 1116–1128 (2006)

Jacob Hinkle earned B.S. degrees in mathematics and physics from Miami University in May 2006. In 2007, he joined the Scientific Computing and Imaging (SCI) Institute at the University of Utah. He received his Ph.D. in Bioengineering from the University of Utah in July 2013. His current research interests include medical image analysis, computational anatomy, and manifold-based statistics.



P. Thomas Fletcher received his B.A. degree in Mathematics at the University of Virginia in 1999. He received an M.S. in Computer Science in 2002 followed by a Ph.D. in Computer Science in 2004 from the University of North Carolina at Chapel Hill. He is currently an Assistant Professor in the School of Computing at the University of Utah. His research interests include manifold statistics, shape analysis, medical image analysis

Sarang Joshi joined SCI as an Associate Professor of the Department of Bio Engineering in 2006. Before coming to Utah, Dr. Joshi was an Assistant Professor of Radiation Oncology and an Adjunct Assistant Professor of Computer Science at the University of North Carolina in Chapel Hill. Prior to joining Chapel Hill Dr. Joshi was Director of Technology Development at IntellX, a Medical Imaging start-up company which was later acquired by Medtronic. Sarang's research interests are in the field of Computational Anatomy. The principal aim of computational anatomy is the development of specialized mathematical and computational tools for the precise study of anatomical variability and the application of these tools for improved medical treatment, diagnosis and understanding of disease. In 2005 he spent a year on sabbatical at DKFZ (German Cancer Research Center) in Heidelberg, Germany, as a visiting scientist in the Department of Medical Physics where he focused on developing four dimensional radiation therapy approaches for improved treatment of Prostate and Lung Cancer. He was also one of the founding partner of Morphormics, Inc. which was recently acquired by Accuray. He has won numerous awards including the 2007 David Marr Best Paper Award, The international journal Signal Processing 2010 Most cited paper Award, and MICCAI 2010 Best of the Journal Issue Award. He holds numerous patents in the area of image registration and segmentation

CHAPTER 5

IRROTATIONAL DIFFEOMORPHISMS

Over the last decade, the field of computational anatomy has substantially matured and several approaches have been developed for the study of anatomical variations that are evident within medical images. The most theoretically developed and principled approaches are based on the Riemannian geometry of groups of diffeomorphisms of three-dimensional Euclidian space, $\mathbb{R}^3$, and its submanifolds (points, curves, and surfaces) on which these groups act. Fundamental to this approach is the computation of geodesics which provide *normal coordinates* via the Riemannian log and exponential maps allowing for statistical analysis of anatomical variability. Despite the elegance of the theory, universal adoption has been limited by the computational complexity of the resulting optimization problems, especially the need for infinite dimensional optimization to compute the geodesic and the log map. To mitigate the computational complexity, recently, some [8] have suggested abandoning the intrinsic Riemannian geometric approach and taking an extrinsic Eulerian view of deformation based on stationary vector fields.

The major contribution of this chapter is the use of a recent result by Modin [10] concerning the polar factorisation of diffeomorphisms (analogous to the polar factorisation of matrices) to define a submanifold of irrotational diffeomorphisms which I call $\mathrm{IDiff}(\mathbb{R}^d)$.[1] In this chapter, I show that using the natural Laplacian metric, this submanifold is flat, meaning that sectional curvature in every direction is zero. This theoretical result has far reaching consequences: for example, within this space, the intrinsic or Fréchet mean is guaranteed to be unique. Another consequence of this remarkable result is that I am able to derive in closed form the Riemannian log map and compute the distance between the identity and any diffeomorphism within $\mathrm{IDiff}(\Omega)$ in closed form. I begin to explore

[1] As noted by Modin, this polar factorisation is related to that of Brenier [4], but uses a different metric and fiber bundle structure. As Modin uses a right invariant metric on diffeomorphisms, his polar factorisation applies only to diffeomorphisms, whereas Brenier's factorisation holds for all vector-valued functions of sufficient smoothness.

the applications of this by developing extremely computationally efficient and numerically stable image registration algorithms.

This work was initiated with the publication of Hinkle and Joshi [5], in which the definition of $\mathrm{IDiff}(\Omega)$ was slightly different. The definition given in this chapter is a restriction of that in [5], but this change has little practical effect, and the algorithms and results are unchanged.

5.1 Mathematical Background and Notation

Although diffeomorphisms in the context of image registration have been extensively studied, for completeness, I review the basic setup. A compactly supported diffeomorphism φ is a bijective map from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that both φ and its inverse φ^{-1} are smooth and have compact support (meaning they are equal to the identity outside a bounded region). The identity transformation id is a diffeomorphism as well as the composition of any two diffeomorphisms. As the inverse of a diffeomorphism is also a diffeomorphism, the set of all diffeomorphisms forms a group. The Lie algebra, $\mathfrak{g}$, of the compactly supported diffeomorphism group, $\mathrm{Diff}_c(\mathbb{R}^d)$, consists of all compactly supported smooth vector fields on $\mathbb{R}^d$, equipped with a Lie bracket of given by the (negative) Jacobi-Lie bracket of vector fields, defined by

$$\mathrm{ad}_v\, w = -[v, w] = (Dv)w - (Dw)v, \tag{5.1}$$

where Dv is the Jacobian matrix of the vector field v.

Given a time-dependent vector field $v(x, t)$ one defines a path in $\mathrm{Diff}_c(\mathbb{R}^d)$ via the ordinary differential equation

$$\frac{d\varphi(x,t)}{dt} = v(\varphi(x,t),t), \text{ with initial condition: } \varphi(x,0) = x.$$

One imposes a right invariant Riemannian metric on $\mathrm{Diff}_c(\mathbb{R}^d)$ by choosing a symmetric (with respect to the L^2 inner product), positive-definite differential operator L which acts on velocity fields. The operator L then determines the norm of a velocity field,

$$\|v\|_{\mathfrak{g}}^2 = \int (Lv(x), v(x))dx. \tag{5.2}$$

The dual space of the Lie algebra, $\mathfrak{g}^*$ consists of vector-valued distributions. The velocity, $v \in \mathfrak{g}$, maps to its dual deformation momenta, $m \in \mathfrak{g}^*$, via the operator L such that $m = Lv$. Using this norm, geodesics are defined as energy minimizing paths between their endpoints.

The distance between the identity and a diffeomorphism ϕ is defined via the minimization problem:

$$d(id,\phi)^2 = \inf\left\{\int_0^1 \|v(\cdot,t)\|_{\mathfrak{g}}^2 dt,\ \text{subject to: } \varphi(\cdot,1)=\phi\right\}. \tag{5.3}$$

5.1.1 EPDiff for Geodesic Evolution

Given the initial velocity, $v_0 \in \mathfrak{g}$, or equivalently, the initial momentum, $m(0) = m_0 \in \mathfrak{g}^*$, the geodesic path $\varphi(t)$ satisfies the EPDiff equation [1, 9]:

$$\frac{d}{dt}m = -\operatorname{ad}_v^* m = -(Dv)^T m - Dmv - (\nabla\cdot v)m \tag{5.4}$$

where D denotes the Jacobian matrix, and the operator ad_v^* is the dual of ad_v [9, 1, 15].

5.2 Polar Factorisation of Diffeomorphisms and $\mathrm{IDiff}(\mathbb{R}^d)$: the Space of Irrotational Diffeomorphisms

Let Ω denote the image domain. In this section, I will define the space $\mathrm{IDiff}(\Omega)$ as what Modin [10] calls the *polar cone* of $\mathrm{Diff}_c(\Omega)$. In order to properly introduce this space, first consider that $\mathrm{Diff}_c(\Omega)$ has a fiber bundle structure in which the base space is that of scalar *densities* and diffeomorphisms project to densities via the Jacobian determinant [10, Eq. 11]:

$$\pi_{vol} : \mathrm{Diff}_c(\Omega) \to \mathrm{Dens}(\Omega) \tag{5.5}$$

$$\pi_{vol}(\varphi) = |D\varphi| \tag{5.6}$$

where $\mathrm{Dens}(\Omega) = \{\nu \in L^1(\Omega) : \nu > 0, \int \nu = 1\}$. Note that Modin was mostly interested in the case when the image domain Ω is compact, while I instead work with compactly-supported functions over a noncompact domain $\Omega = \mathbb{R}^d$. Notice that π_{vol} is invariant under composition by any measure-preserving diffeomorphism $S \in \mathrm{SDiff}(\Omega)$, while any compressible diffeomorphism necessarily alters the projected density. Thus the fibers of $\mathrm{Diff}_c(\Omega)$ are each diffeomorphic to $\mathrm{SDiff}(\Omega)$.

Given a distance metric $d(\cdot,\cdot)$ on $\mathrm{Diff}_c(\Omega)$, the *polar cone* (which I call $\mathrm{IDiff}(\Omega)$ of $\mathrm{Diff}_c(\Omega)$ is then defined as those diffeomorphisms whose closest point within the identity fiber is the identity itself [10, p. 23]:

$$\mathrm{IDiff}(\Omega) = \{\varphi \in \mathrm{Diff}_c(\Omega) : d(id,\varphi) \le d(\phi,\varphi), \forall \phi \in \mathrm{SDiff}(\Omega)\} \tag{5.7}$$

Modin's polar factorisation [4, Thm. 5.6] states that any diffeomorphism $\varphi \in \mathrm{Diff}_c(\Omega)$ can be uniquely written as a composition

$$\varphi = S \circ \psi, \quad \text{where } S \in \mathrm{SDiff}(\Omega), \psi \in \mathrm{IDiff}(\Omega). \tag{5.8}$$

This factorisation is analogous to the classical polar factorisation of matrices, which states that any invertible real matrix $A \in \mathrm{GL}(n)$ can be written as product of an orthogonal matrix $O \in \mathrm{O}(n)$ and a symmetric, positive definite matrix $P \in \mathcal{P}(n)$:

$$A = OP. \tag{5.9}$$

Note that $\mathrm{O}(n)$ forms a well-understood Lie group (analogous to $\mathrm{SDiff}(\Omega)$. On the other hand, $\mathcal{P}(n)$ is not a Lie group, but rather a symmetric space acted upon by other groups. To see this, note that the product of two symmetric matrices need not be symmetric. This is indeed the case for $\mathrm{IDiff}(\Omega)$ as well, which does not itself form a group, since the composition of two diffeomorphisms in $\mathrm{IDiff}(\Omega)$ need not itself Lie in $\mathrm{IDiff}(\Omega)$.

Modin's polar factorisation of $\mathrm{Diff}_c(\Omega)$ is also intimately connected to the Helmholtz-Hodge decomposition of vector fields, which has already proven useful for modeling incompressible deformation in computational anatomy [6, 8] and other fields [3]. The Helmholtz-Hodge decomposition states that any square-integrable vector field $v \in \mathfrak{g}$ can be written as

$$v = \nabla f + \nabla \times A \tag{5.10}$$

where f is a scalar function and A is a vector field. This constitutes a decomposition of $\mathfrak{g}$ into two linear subspaces: one containing irrotational vector fields represented as gradients of scalar Sobolev functions and one containing incompressible (divergence-free) vector fields represented as curls of Sobolev vector fields. I will denote by $\mathfrak{g}_P$ the subspace of irrotational vector fields and by $\mathfrak{g}_S$ the subspace of divergence-free vector fields, so that $\mathfrak{g} = \mathfrak{g}_P \oplus \mathfrak{g}_S$. Assuming compact support, f is uniquely determined by the divergence of v and f satisfies Poisson's equation:

$$\nabla \cdot (v) = g, \quad \Delta f = g, \tag{5.11}$$

where Δ is the Laplacian operator and $g \in L^2(\Omega)$.

$\mathrm{SDiff}(\Omega)$ is commonly defined as all flows from the identity along incompressible vector fields. This definition implies trivially that $\mathrm{SDiff}(\Omega)$ is a group, and since $\mathfrak{g}_S$ is a Lie subalgebra (closed under Lie brackets) $\mathrm{SDiff}(\Omega)$ is itself a group [1]. However, $\mathfrak{g}_P$ is *not* closed with respect to Lie bracket, and so attempts to define another component via all flows along irrotational vector fields lead to sets which may even fail to be manifolds. Instead, our definition chooses a particular point (the identity), and defines $\mathrm{IDiff}(\Omega)$ with respect to this special point and a choice of metric on $\mathrm{Diff}_c(\Omega)$. Modin showed [10,

Thm. 5.6] that $\mathfrak{g}_P$ corresponds precisely to the right invariant horizontal distribution for the fiber structure on $\mathrm{Diff}_c(\Omega)$. Furthermore, he shows that geodesics from the identity in these horizontal directions remain horizontal, and, most importantly, that this leads to an equivalent definition of $\mathrm{IDiff}(\Omega)$ as the set of all geodesics from the identity in irrotational directions:

$$\mathrm{IDiff}(\Omega) = \{\mathrm{Exp}_{id}(\nabla f) : f \in H^1(\Omega)\}. \tag{5.12}$$

5.3 Metric and Geodesics on $\mathrm{IDiff}(\Omega)$

The Helmholtz-Hodge decomposition was mentioned earlier in the context of the polar factorisation of $\mathrm{Diff}_c(\Omega)$. It is fundamental to our purpose in the sense that the Helmholtz-Hodge decomposition describes an infinitesimal version of the polar factorisation with which I have defined $\mathrm{IDiff}(\Omega)$. In this section, I will present the negative Laplacian metric, in this context also called the $\dot{H}^1$ metric, which I claim is natural with respect to the Helmholtz-Hodge decomposition.

First, consider the negative vector Laplacian metric

$$\langle v, w\rangle_{\mathfrak{g}} = \int_\Omega (-\Delta v(x))^T w(x) dx. \tag{5.13}$$

Although the Laplacian operator Δ has a null space containing linear functions, due to the compact support constraint on $\mathfrak{g}$, this null space is excluded, making $-\Delta$ a positively defined linear operator on $\mathfrak{g}$. Hence, the above is a valid metric on $\mathfrak{g}$ with which I define a right invariant metric on $\mathrm{Diff}_c(\Omega)$ (and hence on $\mathrm{IDiff}(\Omega)$).

Now consider the following identity (Lagrange's formula):

$$\Delta v = \nabla \times \nabla \times v - \nabla\nabla \cdot v. \tag{5.14}$$

Combining this with the Helmholtz-Hodge decomposition and the identities

$$\nabla \times \nabla f = 0 \tag{5.15}$$

$$\nabla \cdot \nabla \times A = 0, \tag{5.16}$$

I rewrite the Laplacian metric as

$$\langle v, v\rangle_{\mathfrak{g}} = \langle \nabla f + \nabla \times A, \nabla f + \nabla \times A\rangle_{\mathfrak{g}} \tag{5.17}$$
$$= \int_\Omega \left(\nabla \times \nabla \times \nabla \times A - \nabla\nabla \cdot \nabla f\right)^T v dx \tag{5.18}$$
$$= \int_\Omega \left(\nabla \times \nabla \times \nabla \times A\right)^T \left(\nabla f + \nabla \times A\right) - \nabla\left(\nabla \cdot \nabla f\right)^T \left(\nabla f + \nabla \times A\right) dx \tag{5.19}$$
$$= \int_\Omega \left(\nabla \times \nabla \times A\right)^T \nabla \times \left(\nabla f + \nabla \times A\right) + \left(\nabla \cdot \nabla f\right) \nabla \cdot \left(\nabla f + \nabla \times A\right) dx \tag{5.20}$$
$$= \|\nabla \times \nabla \times A\|^2_{L^2(\Omega)} + \|\nabla \cdot \nabla f\|^2_{L^2(\Omega)}. \tag{5.21}$$

This means that the Laplacian metric is the L^2 metric of the curl of $\mathfrak{g}_S$ and of the divergence on $\mathfrak{g}_P$:

$$\langle v, w\rangle_{\mathfrak{g}_S} = \langle \nabla \times v, \nabla \times w\rangle_{L^2(\Omega,\mathbb{R}^3)} \tag{5.22}$$
$$\langle v, w\rangle_{\mathfrak{g}_P} = \langle \nabla \cdot v, \nabla \cdot w\rangle_{L^2(\Omega,\mathbb{R})} \tag{5.23}$$

Note that the formula for the norm is derived above, but the inner product exhibits a very similar splitting into $\mathfrak{g}_S$ and $\mathfrak{g}_P$ terms. This shows that the Laplacian metric, which I will employ for the remainder of this chapter, is naturally block diagonal with blocks described by the Helmholtz-Hodge decomposition.

Now, examining the restriction of the Laplacian metric to $\mathfrak{g}_P$, we can write the metric in terms of the scalar function f:

$$\langle v, w\rangle_{\mathfrak{g}_P} = \langle \nabla \cdot v, \nabla \cdot w\rangle_{L^2} = \langle \nabla \cdot \nabla f, \nabla \cdot \nabla h\rangle_{L^2} = \int_\Omega \Delta f(x) \Delta h(x) dx \tag{5.24}$$

where Δ is the scalar Laplacian operator. Notice that with the above inner product, if g is the divergence of v, the norm of v is simply the L^2 norm of g:

$$\|v\|^2_{\mathfrak{g}_P} = \|\nabla \cdot v\|^2_{L^2(\Omega)} = -\int (\nabla\nabla \cdot v(x))^T v(x) dx = \|g\|^2_{L^2(\Omega)}. \tag{5.25}$$

Letting $\psi \in \mathrm{IDiff}(\Omega)$ be an irrotational diffeomorphism, a geodesic between ψ and the identity is a path $\alpha(t) \in \mathrm{IDiff}(\Omega)$ connecting ψ and the identity that minimizes

$$S(\alpha) = \frac{1}{2}\int_0^1 \|\dot\alpha(t)\|^2 dt = \frac{1}{2}\int_0^1 \int |(\Delta f(t))(x)|^2 dx dt. \tag{5.26}$$

Geodesics on $\mathrm{IDiff}(\Omega)$ passing through the identity are actually minimizing curves in all of $\mathrm{Diff}_c(\Omega)$ with the constraint that the right-trivialized velocity lie in $\mathfrak{g}_P$ [10].

I define the momentum associated with the velocity v as $m = -\nabla\nabla\cdot v = -\nabla g$. Geodesics in $\mathrm{IDiff}(\Omega)$ satisfy the Euler-Poincaré equation, (5.4), with the constraint that v is curl free. Substituting $m = -\nabla g, v = \nabla f$, and $\nabla\cdot v = g$, the Euler-Poincaré equation simply becomes:

$$\frac{d}{dt}\nabla g = -Hg\nabla f - (Hf)^T\nabla g - g\nabla g. \tag{5.27}$$

The Hessian matrix is always symmetric and notice that $\nabla(g^2) = 2g\nabla g$, so we can rewrite this using the product rule as

$$\nabla\frac{d}{dt}g = -\nabla\left(\nabla g^T\nabla f + \frac{1}{2}g^2\right). \tag{5.28}$$

Along with our boundary conditions on g, this implies that

$$\dot{g} + \nabla g^T v = -\frac{1}{2}g^2. \tag{5.29}$$

The left-hand side has the form of a material derivative, suggesting a change to Lagrangian coordinates. Introducing $\gamma(t) = g\circ\psi(t)$, implying $\dot{\gamma} = \dot{g}\circ\psi + ((\nabla g)^T v)\circ\psi$ we see that

$$\dot{\gamma}(t) = -\frac{1}{2}\gamma(t)^2, \quad \text{or} \quad \gamma(t) = \frac{\gamma(0)}{\frac{1}{2}t\gamma(0)+1} \quad . \tag{5.30}$$

Using the shorthand $g_0 = g(0)$ and the assumption $\psi(0) = id$, we arrive at

$$g(t)\circ\psi(t) = \frac{g_0}{\frac{1}{2}tg_0+1}. \tag{5.31}$$

The quantity $g(t)$ is, by definition, the divergence of the velocity at time t. Using the well-known Liouville's formula, I relate this directly to the determinant of the Jacobian matrix of the diffeomorphism ψ as follows:

$$|D\psi(t)| = \exp\int_0^t(\nabla\cdot v)\circ\psi ds = \exp\int_0^t\frac{g_0}{\frac{1}{2}sg_0+1}ds = \left(\frac{1}{2}tg_0+1\right)^2. \tag{5.32}$$

Using the solution of the EPDiff equation, we can explicitly write the expression for the distance in $\mathrm{IDiff}(\Omega)$ between the identity and any irrotational diffeomorphism ψ. As the metric is simply the L^2 norm of g, by conservation of momenta along a geodesic, we have

$$d(id,\psi)^2 = \|g_0\|^2_{L^2(\Omega)} = 4\int_{\mathbb{R}^d}(\sqrt{|D\psi|}-1)^2dx. \tag{5.33}$$

The simplicity of the above formula comes from the fact that by solving the EPDiff equation, g_0 is essentially the log map on $\mathrm{IDiff}(\Omega)$ with the $\dot{H}^1$ metric.

5.4 Curvature of $\mathrm{IDiff}(\Omega)$

We now use the relationship between g_0 and $|D\psi|$ to show that the curvature of $\mathrm{IDiff}(\Omega)$ with the $\dot{H}^1$ metric is 0. Define the following mapping from ψ to the divergence of its initial velocity field:

$$P : \mathrm{IDiff}(\Omega) \to L^2(\Omega) \tag{5.34}$$

$$P(\psi) = 2(\sqrt{|D\psi|} - 1) = g_0. \tag{5.35}$$

Notice that this function is defined on all of $\mathrm{Diff}_c(\Omega)$, but is unaffected by the incompressible component of ψ.

Lemma 5.1 *The pushforward of a vector field $u \circ \psi \in T_\psi \mathrm{IDiff}(\Omega)$ under the mapping P is given by the formula*

$$TP(u \circ \psi) = \sqrt{|D\psi|}(\nabla \cdot u) \circ \psi. \tag{5.36}$$

Proof : Let ψ_s be a family of irrotational diffeomorphisms indexed by the real variable s and satisfying

$$\psi_0 = \psi, \quad \frac{d}{ds}|_{s=0}\psi_s = u \circ \psi. \tag{5.37}$$

Then the pushforward of the vector field u is defined as

$$TP(u \circ \psi) = \frac{d}{ds}|_{s=0} P\psi_s. \tag{5.38}$$

A straightforward computation then yields

$$TP(u \circ \psi) = 2\frac{d}{ds}|_{s=0}\sqrt{|D\psi_s|} = \sqrt{|D\psi|}(\nabla \cdot u) \circ \psi. \tag{5.39}$$

Theorem 5.1 *The mapping P is an isometry from $\mathrm{IDiff}(\Omega)$ into an open subset of $L^2(\Omega)$.*

Proof : As the pushfoward is only zero for divergence-free vector fields, Lemma 5.1 directly implies that P is injective on $\mathrm{IDiff}(\Omega)$. To prove that P is furthermore an isometry, I compute the pullback of the L^2 metric for any two vector fields $u \circ \psi, w \circ \psi \in T_\psi \mathrm{IDiff}(\Omega)$:

$$\langle u, w \rangle_{P^*} = \langle TP(u \circ \psi), TP(w \circ \psi) \rangle_{L^2(\Omega)}. \tag{5.40}$$

Plugging in and performing a change of variables, we have

$$\langle u, w \rangle_{P^*} = \int \sqrt{|D\psi(x)|}(\nabla \cdot u) \circ \psi(x) \sqrt{|D\psi(x)|}(\nabla \cdot w) \circ \psi(x) dx \tag{5.41}$$

$$= \int |D\psi(x)|(\nabla \cdot u) \circ \psi(x)(\nabla \cdot w) \circ \psi(x) dx \tag{5.42}$$

$$= \langle \nabla \cdot u, \nabla \cdot v \rangle_{L^2(\Omega)}, \tag{5.43}$$

which is our right invariant metric on $\mathrm{IDiff}(\Omega)$, proving that P is a local isometry. By the uniqueness of Modin's polar factorisation, the mapping P is injective, completing the proof. $\square$

The property that P is an isometry is remarkable in that it implies (since $L^2(\Omega)$ is a flat vector space) that with the $\dot{H}^1$ metric, $\mathrm{IDiff}(\Omega)$ has zero Riemannian curvature[2]. Another important consequence is that under P, geodesics in $\mathrm{IDiff}(\Omega)$ map to straight lines in $L^2(\Omega)$. The image of P consists of all L^2 functions with values strictly greater than -2, implying that geodesics can leave this open subset in finite time. Given an initial velocity field, this blow-up time is determined by the minimum value of its divergence g_0 and (5.32).

The P map is injective, so given $g_0 \in L^2(\Omega)$, there is a unique irrotational diffeomorphism $\psi \in \mathrm{IDiff}(\Omega)$ in the inverse image $P^{-1}(g_0)$. Computation of ψ is equivalent to computing the exponential map in $\mathrm{IDiff}(\Omega)$. I am unaware of a closed-form method for computing ψ, but it may be computed numerically using (5.31) to compute $g(t) = \nabla \cdot v(t)$ at each time, then solving for the velocity field $v(t)$ and integrating the flow.

5.5 Irrotational Image Registration

Consider a registration problem in which two images $I_0, I_1 \in L^2(\Omega)$ are given and one wishes to find an irrotational deformation $\psi \in \mathrm{IDiff}(\Omega)$ that best matches the two images. Analogous to the LDDMM approach, I introduce the energy functional

$$E(\psi) = \frac{1}{2\sigma^2}\|I_0 \circ \psi^{-1} - I_1\|^2_{L^2(\Omega)} + d(id, \psi)^2 \tag{5.44}$$

where d denotes the geodesic distance within $\mathrm{IDiff}(\Omega)$. However, unlike with general LDDMM, the distance term can now be evaluated in closed form only using ψ:

$$E(\psi) = \frac{1}{2\sigma^2}\|I_0 \circ \psi^{-1} - I_1\|^2_{L^2(\Omega)} + 4\|\sqrt{|D\psi|} - 1\|^2_{L^2(\Omega)}. \tag{5.45}$$

This allows us to take the Sobolev variation of E with respect to ψ directly by first taking the L^2 variation and then sharping it using the inverse of the metric. Let $\nabla c \in \mathfrak{g}_P$ be a perturbation of ψ, and let $\psi_s \in \mathrm{IDiff}(\Omega)$ be a family of irrotational diffeomorphisms parametrized by the real variable s, satisfying

$$\psi_0 = \psi \quad \text{and} \quad \frac{d}{ds}|_{s=0}\psi_s = (\nabla c) \circ \psi. \tag{5.46}$$

[2]This has been observed very recently in [2] for the special case of $d = 1$ where $\mathrm{IDiff}(\mathbb{R}^1) = \mathrm{Diff}_c(\mathbb{R}^1)$ as the only compactly-supported measure-preserving diffeomorphism of the real line is the identity mapping.

Then the variation of E with respect to ψ in the direction of ∇c is computed via

$$(\delta E, \nabla c) = \frac{d}{ds}|_{s=0} E(\psi_s) \tag{5.47}$$

$$= \frac{d}{ds}|_{s=0} \frac{1}{2\sigma^2} \int_\Omega (I_0 \circ \psi_s^{-1}(y) - I_1(y))^2 dy + 4 \int_\Omega (\sqrt{|D\psi_s(x)|} - 1)^2 dx \tag{5.48}$$

$$= \frac{1}{\sigma^2} \int_\Omega (I_0 \circ \psi^{-1}(y) - I_1(y)) \nabla (I_0 \circ \psi^{-1}(y))^T \nabla c(y) dy \tag{5.49}$$

$$+ 4 \int_\Omega (\sqrt{|D\psi(x)|} - 1) \frac{1}{\sqrt{|D\psi(x)|}} \frac{d}{ds}|_{s=0} |D\psi_s(x)| dx. \tag{5.50}$$

Using $\frac{d}{ds}|_{s=0}|D\psi_s(x)| = (\nabla \cdot \nabla c) \circ \psi(x)|D\psi(x)|$ and the fact that, for compactly supported vector fields, the adjoint of the divergence is the negative gradient, we have

$$(\delta E, \nabla c) = -\frac{1}{\sigma^2} \int_\Omega \nabla \cdot \left((I_0 \circ \psi^{-1}(y) - I_1(y)) \nabla (I \circ \psi^{-1}(y)) \right) c(y) dy \tag{5.51}$$

$$+ 4 \int_\Omega (\sqrt{|D\psi| \circ \psi^{-1}(y)} - 1) \sqrt{|D\psi| \circ \psi^{-1}(y)} \Delta c(y) |D\psi^{-1}(y)| dy. \tag{5.52}$$

Now use the identity $(D\psi^{-1}) \circ \psi(x) = (D\psi)^{-1}(x)$ and self-adjointness of the Laplacian to simplify this to

$$(\delta E, \nabla c) = -\frac{1}{\sigma^2} \int_\Omega \nabla \cdot \left((I_0 \circ \psi^{-1}(y) - I_1(y)) \nabla (I_0 \circ \psi^{-1}(y)) \right) c(y) dy \tag{5.53}$$

$$+ 4 \int_\Omega c(y) \Delta \left(1 - \sqrt{|D\psi^{-1}(y)|} \right) dy. \tag{5.54}$$

By adjointing the gradient in the left-hand side, we see that since this must hold for all c, we have

$$\nabla \cdot \delta E = \frac{1}{\sigma^2} \nabla \cdot \left((I_0 \circ \psi^{-1} - I_1) \nabla (I_0 \circ \psi^{-1}) \right) + 4\Delta (\sqrt{|D\psi^{-1}|} - 1). \tag{5.55}$$

In order to convert δE to the Sobolev variation of E, solve the following for the scalar function b:

$$\Delta^2 b = \frac{1}{\sigma^2} \nabla \cdot \left((I_0 \circ \psi^{-1} - I_1) \nabla (I_0 \circ \psi^{-1}) \right) + 4\Delta (\sqrt{|D\psi^{-1}|} - 1) \tag{5.56}$$

then update ψ via $\psi(x) \mapsto \psi(x) - \epsilon(\nabla b) \circ \psi(x)$ for some step-size ϵ. In practice, as ψ is never needed, I directly update only ψ^{-1} via $\psi^{-1}(y) \mapsto \psi^{-1}(y + \epsilon \nabla b(y))$.

Notice that this allows ψ^{-1} to be optimized directly in a gradient-based scheme without the need for numeric integration of geodesic equations or adjoint equations.

5.6 Symmetric Image Registration

In this section, I present an image registration approach that is symmetric with respect to swapping of the input images. Consider re-weighting the image match term by the square root of the Jacobian determinant.

$$E(\psi) = \frac{1}{2\sigma^2} \int |I_0 \circ \psi^{-1}(y) - I_1(y)|^2 \sqrt{|D\psi^{-1}(y)|} dy + d(id, \psi)^2. \tag{5.57}$$

Now using the change of variables $x = \psi^{-1}(y)$

$$E(\psi) = \frac{1}{2\sigma^2} \int \|I_0(x) - I_1 \circ \psi(x)\|^2 \sqrt{|(D\psi^{-1}) \circ \psi(x)|} |D\psi(x)| dx + d(\psi^{-1}, id)^2. \tag{5.58}$$

Using the inversion-invariance of our metric, we rewrite the cost functional as

$$E(\psi) = \frac{1}{2\sigma^2} \int |I_0(x) - I_1 \circ \psi(x)|^2 \sqrt{|D\psi(x)|} dx + d(id, \psi)^2. \tag{5.59}$$

This has the same form as the original function in which the first image I_0 was deformed to match I_1, but instead, we match I_1 to I_0. So the introduction of the square-root Jacobian determinant into the image match term has the effect of making the image registration problem invariant under relabeling of the input images. This resembles the "square-root trick" used in one dimension to develop parametrization-invariant metrics on time-series data [11] and planar curves [13].

Computing the variation of this functional is very similar to the method in the previous section, and leads us to the following biharmonic equation:

$$\Delta^2 b = \frac{1}{\sigma^2} \nabla \cdot \left((I_0 \circ \psi^{-1} - I_1) \sqrt{|D\psi^{-1}|} \nabla (I_0 \circ \psi^{-1}) \right) \tag{5.60}$$

$$+ \Delta \left(\left(4(1 - \sqrt{|D\psi^{-1}|}) - \frac{1}{4\sigma^2} |I_0 \circ \psi^{-1} - I_1|^2 \right) \sqrt{|D\psi^{-1}|} \right). \tag{5.61}$$

After solving this equation for b, we take the gradient then update ψ just as we did in the asymmetric case.

5.6.1 Neuroimaging Study

I have implemented the symmetric irrotational image registration algorithm and applied it to two structural MRI images. Figure 5.1 shows the result of symmetric irrotational image registration. Notice that even without allowing any local rotation, the two images are matched quite well. In the bottom row is shown the energy at each iteration, indicating very stable convergence. Also notice that the Jacobian determinant clearly indicating regions of expansion and contraction. In our irrotational matching method, the Jacobian determinant entirely characterizes the diffeomorphism.

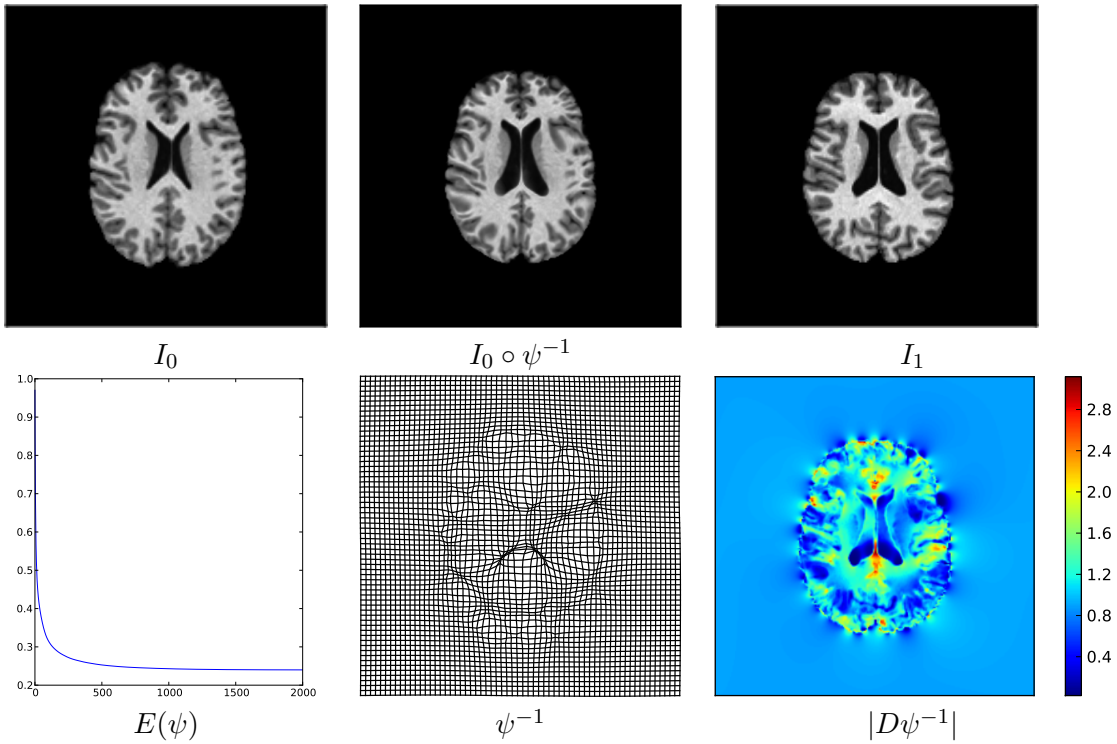


Figure 5.1. Neuroimaging study, symmetric irrotational registration results. The algorithm was run with inputs I_0, I_1 to generate the irrotational diffeomorphism ψ. The plot of energy $E(\psi)$ at each iteration is shown on the left in the lower column showing good convergence, along with the estimated deformation ψ and its Jacobian determinant.

5.7 Hybrid Irrotational/Incompressible Registration

The irrotational registration algorithms presented so far are clearly useful, but there are occasions when a significant rotational component is necessary to achieve good matching. In this section, I present an extension of the irrotational-only algorithms which allows an incompressible component to be estimated without any penalty. The resulting algorithm provides fully diffeomorphic deformations while retaining the efficiency of the purely irrotational version.

Consider registration using a general diffeomorphism $\varphi = S \circ \psi$, where $S \in \mathrm{SDiff}(\Omega)$. Using the polar factorisation of $\mathrm{Diff}_c(\Omega)$, I replace $E(\psi)$ with the functional

$$E(\psi, S) = \frac{1}{2\sigma^2} \int |I_0 \circ \psi^{-1} \circ S^{-1} - I_1|^2 \sqrt{|D\psi^{-1} \circ S^{-1}|} dy + d(id, \psi)^2. \tag{5.62}$$

Equation 5.62 is easily rewritten in terms of φ^{-1} only, using a simple change of variables and the fact that $|DS| = 1$:

$$E(\varphi) = \frac{1}{2\sigma^2} \int |I_0 \circ \varphi^{-1} - I_1|^2 \sqrt{|D\varphi^{-1}|} dy + 4 \int (\sqrt{|D\varphi^{-1}|} - 1)^2 dy. \tag{5.63}$$

This is optimized by decomposing the Sobolev variation of $E(\varphi)$ using the Helmholtz-Hodge decomposition into irrotational and incompressible components, then performing gradient descent steps in either component. The irrotational updates are performed exactly as described in the previous section, and since the incompressible updates do not effect the Jacobian determinant, the incompressible update direction $w \in \mathfrak{g}_S$ is found by simply solving

$$\Delta w = -\frac{1}{\sigma^2}\left(I_0 \circ \varphi^{-1} - I_1\right)\sqrt{|D\varphi^{-1}|}\nabla(I_0 \circ \varphi^{-1}) \tag{5.64}$$

and projecting onto the space of divergence-free vector fields $\mathfrak{g}_S$. This projection has been discussed previously in the literature and is performed efficiently in the Fourier domain while simultaneously solving the above Poisson's equation [6].

5.7.1 Synthetic Example

In order to test the performance of our algorithms in the presence of large deformation, a simulated experiment was also performed. Two synthetic two-dimensional datasets were generated, simulating a completed "C" and a half C. In Fig. 5.2 are shown the results of the hybrid image registration algorithm. Notice that the deformed half C image, $I_0 \circ \varphi^{-1}$, agrees very well with the full C image, I_1, and that this is achieved while maintaining a diffeomorphic transformation. Since I penalize the L^2 norm of the square root Jacobian, the Jacobian determinant of the overall deformation is distributed very evenly across the entire deforming region, instead of being concentrated at a single advancing edge.

5.8 Atlas-Building

The hybrid image matching algorithm described previously is trivially adapted to an atlas construction method resembling that of Joshi et al. [7]. Given a collection of images J_j, an atlas image I_0 is defined as an image minimizing the following sum of terms resembling (5.45):

$$E(I_0) = \sum_j \min_{\varphi_j} \frac{1}{2\sigma^2}\|I_0 \circ \varphi_j^{-1} - I_1\|^2_{L^2(\Omega)} + 4\|\sqrt{|D\varphi_j|} - 1\|^2_{L^2(\Omega)}. \tag{5.65}$$

where the deformations φ_j may be considered *residual* deformations whose irrotational factors are to be minimized. Minimization of (5.65) is performed using a straight-forward alternating optimization scheme. The variation of (5.65) and gradient descent with respect to φ_j is computed in exactly the same manner as in hybrid image registration. However,

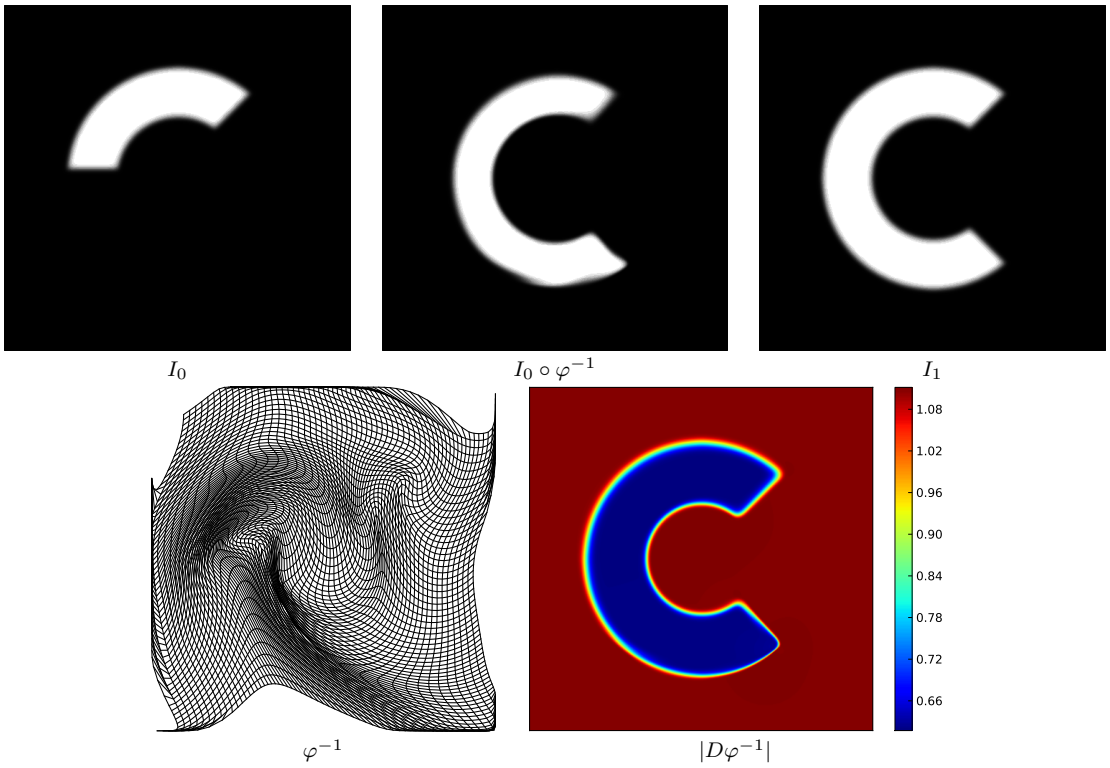


Figure 5.2. Synthetic study, symmetric hybrid image registration results. Shown in the top row are are the input image I_0, along with the deformed image $I_0 \circ \varphi^{-1}$ and the target image I_1. The deformation φ^{-1} and its Jacobian determinant $|D\varphi^{-1}|$ are shown below. The Jacobian determinant is extremely evenly distributed without the C shape.

those gradient descent steps are interleaved with updates to I_0, which, holding all φ_j fixed, minimizes (5.65) when

$$I_0 = \frac{\sum_j |D\varphi_j| J_j \circ \varphi_j}{\sum_j |D\varphi_j|}. \tag{5.66}$$

Note that atlas construction using scalar momentum geodesic shooting requires gradient descent not only on the deformations, but also on the atlas image [14]. The closed form solution for I_0 used above is the same as that for atlas construction using the vector momentum formulation of geodesic shooting recently developed by Singh et al. [12, Eq. 12]. With IDiff(Ω)-based atlas construction, we reap the benefits of this closed-form optimal atlas while also avoiding the timestepping integration involved in geodesic shooting. The result is an extremely efficient algorithm for atlas construction which still uses well-formed geodesic distance penalties.

5.8.1 Atlas Construction Study

Atlas construction was performed on a set of 4T MRI images from 73 subjects. The images are of size 200x225x200 voxels, and the atlas construction algorithm was run with $\sigma = 10$ and a manually tuned gradient descent step size. The resulting image, shown in Fig. 5.3, is quite crisp. Of particular note is the considerable speed which with this atlas was computed. The single-GPU implementation of $\mathrm{IDiff}(\Omega)$-based atlas construction converged in ten minutes, while the equivalent LDDMM atlas construction algorithm requires at least two hours on a cluster of 128 GPUs.

5.9 Discussion

I have shown that Modin's polar factorisation of compactly-supported diffeomorphisms, along with the divergence metric on the irrotational component, leads to novel new image registration algorithms. Furthermore, Thm. 5.1 shows that with this metric, the $\mathrm{IDiff}(\Omega)$ component can be isometrically embedded in the flat vector space $L^2(\Omega)$, a fact that underlies the efficiency of our new algorithms. Even more importantly, it has far reaching statistical implications, allowing statistics to be performed in $\mathrm{IDiff}(\Omega)$ without the difficulties that often accompany statistics on curved manifolds.

Acknowledgements

The authors thank Xavier Pennec and Marco Lorenzi for discussions about irrotational diffeomorphisms, as well as Peter Michor and Martin Bauer for invaluable discussions on the flatness of diffeomorphism spaces on $\mathbb{R}^1$, all of which occurred primarily at the Workshop on Geometry and Statistics at Sonderborg, Denmark in October 2012 organized by the

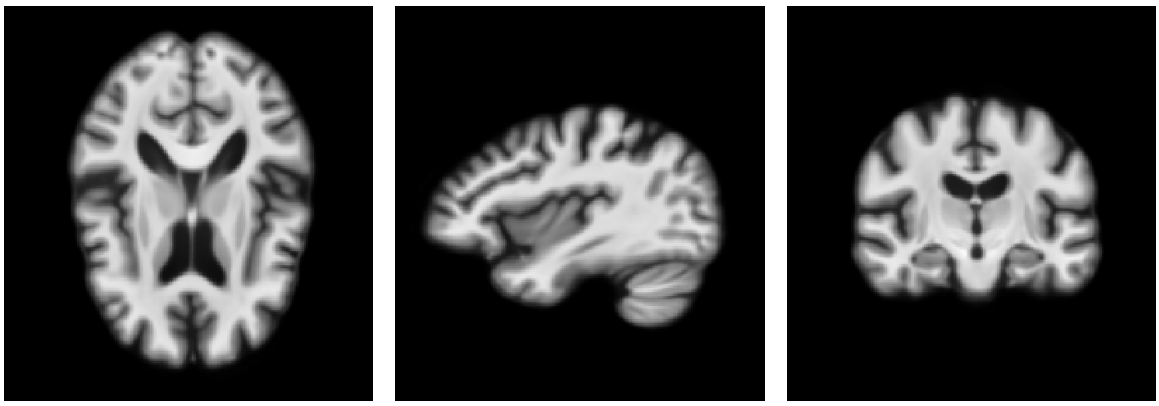


Figure 5.3. Atlas image computed using 73 4T MRI images and hybrid IDiff atlas construction.

University of Copenhagen and Aarhus University. Klas Modin has also been very helpful in recent discussions about IDiff. This work was supported by NIH grants 5R01EB007688, P41 RR023953 and 5R21HL110059-02.

References

[1] Vladimir I Arnol'd. "Sur la Géométrie Différentielle des Groupes de Lie de Dimension Infinie et ses Applications à l'Hydrodynamique des Fluides Parfaits". In: *Ann. Inst. Fourier* 16 (1966), pp. 319–361.

[2] M Bauer, M Bruveris, and P W Michor. "The Homogeneous Sobolev Metric of Order One on Diffeomorphism Groups on the Real Line". In: *arXiv Preprint math-ap/1209.2836* (2012). arXiv: `1209.2836 [math.AP]`.

[3] H. Bhatia et al. "The Helmholtz-Hodge Decomposition: A Survey". In: *IEEE Transactions on Visualization and Computer Graphics* 19.8 (2013), pp. 1386–1404. ISSN: 1077-2626.

[4] Yann Brenier. "Polar Factorization and Monotone Rearrangement of Vector-Valued Functions". In: *Communications on Pure and Applied Mathematics* 44.4 (1991), pp. 375–417.

[5] Jacob Hinkle and Sarang Joshi. "IDiff: Irrotational Diffeomorphisms for Computational Anatomy". In: *Information Processing in Medical Imaging (IPMI)*. Springer. 2013, pp. 754–765.

[6] Jacob Hinkle et al. "4D CT Image Reconstruction with Diffeomorphic Motion Model". In: *Medical Image Analysis* 16.6 (2012), pp. 1307–1316.

[7] Sarang Joshi et al. "Unbiased Diffeomorphic Atlas Construction for Computational Anatomy". In: *NeuroImage* 23 (2004), S151–S160.

[8] M. Lorenzi, N. Ayache, and X. Pennec. "Regional Flux Analysis of Longitudinal Atrophy in Alzheimer's Disease". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012* (2012), pp. 739–746.

[9] Michael I. Miller, Alain Trouvé, and Laurent Younes. "Geodesic Shooting for Computational Anatomy". In: *Journal of Mathematical Imaging and Vision* 24.2 (2006), pp. 209–228. DOI: `10.1007/s10851-005-3624-0`.

[10] K. Modin. "Generalised Hunter-Saxton Equations, Optimal Information Transport, and Factorisation of Diffeomorphisms". In: *ArXiv e-prints* (Mar. 2012). arXiv: `1203.4463 [math-ph]`.

[11] Mauro Piccioni, Sergio Scarlatti, and Alain Trouvé. "A Variational Problem Arising from Speech Recognition". In: *SIAM J. Appl. Math.* 58.3 (June 1998), pp. 753–771.

[12] Nikhil Singh et al. "A Vector Momentum Formulation of Diffeomorphisms for Improved Geodesic Regression and Atlas Construction". In: *International Symposium on Biomedical Imaging (ISBI)*. Apr. 2013.

[13] Anuj Srivastava et al. "Shape Analysis of Elastic Curves in Euclidean Shapes". In: *IEEE Trans. Pattern Anal. and Machine Intel.* 33.7 (July 2011), pp. 1415–1428.

[14] François-Xavier Vialard et al. “Diffeomorphic Atlas Estimation Using Geodesic Shooting on Volumetric Images”. In: *Annals of the BMVA* 5 (2012), pp. 1–12.

[15] L Younes, F Arrate, and M I Miller. “Evolutions Equations in Computational Anatomy”. In: *NeuroImage* 45.1 (2009), S40–S50.

CHAPTER 6

DISCUSSION

This chapter discusses the contributions of this thesis, and presents possibilities for future work.

6.1 Summary of Contributions

The claims presented in the introduction are listed below, with each claim followed by a summary of how it was achieved in previous chapters.

Chapter 2: *A four-dimensional (4D) image reconstruction method is presented, in which a diffeomorphic motion model is used to estimate organ motion using raw projection data while simultaneously estimating a deforming base image. Results are presented for conebeam and fanbeam CT, in phantom studies as well as on patient data, validating the accuracy of the obtained motion estimates. As part of this framework, I derive a method of enforcing an incompressibility constraint, globally or locally, during motion estimation.*

The 4D MAP image econstruction algorithm was shown to provide accurate motion estimates in multiple phantom validation studies. On real data, the algorithm provided a substantial increase in image quality and fewer motion artifacts, as compared to binning methods. In addition, the incompressibility constraint was shown to provide motion estimates with no local expansion or contraction. The spatially-varying incompressibility constraint simultaneously estimates realistic deformations of lung (compressible) and the nearby liver (incompressible). Results were shown using CT data, but extension to other modalities is facilitated by the abstract derivation presented in Chapter 2.

An application of 4D image reconstruction was also presented (the reader is referred to Appendix C). In this radiation oncology application, the motion models obtained using 4D image reconstruction were used to generate stochastic models of dose de-

livery. Using this method, given a static radiation treatment plan, we computed a probabilistic map of dose, computing means and variances of delivered dose at each point in the subject's anatomy.

Chapter 3: *A related application, in which motion estimation is used to correct the pose information during 3D image reconstruction from an uncalibrated imaging device, is presented. Whereas in 4D image reconstruction, the scanner geometry is known and anatomical motion is estimated, in this application, the situation is reversed. The anatomy is presumed to be static, while the scanner geometry is dynamic and not well-calibrated.*

Gradients were derived with respect to intrinsic and extrinsic scanner geometry parameters, and a gradient-based optimization scheme was used along with expectation-maximization maximum likelihood estimation to perform autocalibrating image reconstruction. Scanner geometry estimation was shown to drastically improve image reconstruction. Furthermore, autocalibrating reconstruction was shown to perform well even in cases of limited angular coverage.

Chapter 4: *The established geodesic regression method [4, 10] is extended by introducing a framework for fitting higher-order polynomials on Riemannian manifolds and Lie groups. These more flexible classes of curves enable more accurate curve fitting, while maintaining a compact representation.*

The rolling maps of Jupp and Kent [7] provide insight into this new class of polynomial curves, as discussed at the end of Chapter 4. Riemannian polynomials are defined for any order, including even order polynomials, in contrast to alternative higher-order splines. The form of Riemannian polynomials makes them particularly convenient for use in an adjoint-based shooting optimization scheme. I derived explicit adjoint equations for many special cases, from general manifolds with only a Riemannian structure, to Lie groups with right or left invariant metrics, and finally to data (such as images) given in spaces acted upon by such Lie groups. These cases cover all common settings in computational anatomy, and the solutions exploit the special structure available in each case.

The polynomial adjoint equations for Lie groups and Lie group actions, in particular, are very interesting, as the first adjoint variable λ_0 satisfies an Euler-Poincaré-like

equation[1] whose solution is given by a momentum map. This is an indication that the gradients used in polynomial regression exhibit a considerable amount of *symmetry*. A thorough interpretation of this finding will come with future exploration of polynomial regression in these spaces.

Chapter 5: *A new space of* irrotational *diffeomorphisms, called* IDiff*, is introduced. The geometry of this space enables extremely efficient image registration and atlas-building algorithms. The deformations in* IDiff *are determined chiefly by local expansion and contraction. In light of the conventional study of expansion and contraction in neurodevelopment studies, irrotational diffeomorphisms are potentially a more realistic model for growth and aging of brain structures. In addition, the exploration of this space in real applications gives new insights into the structure of the diffeomorphism group.*

Modin's polar factorisation [8] is used to define $\mathrm{IDiff}(\Omega)$, a space of *irrotational* diffeomorphisms. This polar factorisation is analogous to that of invertible matrices, which splits the space of invertible matrices into a Lie subgroup containing rotation matrices and another space (which is not a group) of symmetric positive-definite matrices. Thus the space $\mathrm{IDiff}(\Omega)$ can be considered an infinite-dimensional analog of symmetric positive-definite matrices.

Although this space is not a Lie group, it nonetheless has surprising and interesting geometric properties. In particular, we give an injective isometric mapping from $\mathrm{IDiff}(\Omega)$ into a convex open subset of $L^2(\Omega)$, whose existence implies that the space $\mathrm{IDiff}(\Omega)$ is geodesically convex and has zero curvature. The full implications of flatness will only be revealed with detailed study of this space, but we have already used it to derive methods for image matching and atlas building whose performance is orders of magnitude better than existing diffeomorphic methods involving geodesic shooting.

6.2 Outlook and Future Work

6.2.1 4D MAP and Autocalibrating Image Reconstruction

We have shown that four-dimensional image reconstruction is practical and gives accurate estimates for respiratory motion during CT image acquisition. Further, the motion artifacts common in binned 4D RCCT imaging make it clear that 4D reconstruction is not only a useful tool, but is *necessary* not only for obtaining motion information, but

[1] Recall that the Euler-Poincaré equation describes the momentum of geodesics in these spaces.

for realistic image reconstruction in the presence of motion. The diffeomorphic motion fields provided by our model give extremely useful information, which will enable treatment planning and beam optimization to benefit by having a more accurate knowledge of tumor shape and location. The publication Geneser et al. [5] (see Appendix C) represents a contribution in this direction, in which the motion fields used in 4D image reconstruction are used to help characterize the expected variability in delivered dose during radiation therapy.

Four-dimensional MAP image reconstruction is a general framework, encompassing applications to multiple imaging modalities and noise models. As shown with the incompressibility constraint, it even allows customization of the motion model. In a previous publication, Hinkle et al. [6], I have shown a proof of concept application of 4D MAP image reconstruction to Fourier MRI data. Indeed, a clear avenue for future development of 4D MAP image reconstruction is the exploration of new application domains and imaging modalities.

Autocalibrating CT reconstruction is interesting from a theoretical point of view because it is the counterpart to 4D image reconstruction; instead of having well-calibrated scanner geometry and unknown subject motion, the subject motion is fixed and scanner geometry is estimated. Currently, only a maximum-likelihood estimate of scanner geometry is computed, but it may be necessary in some cases to impose temporal smoothness priors on the geometry parameters. The resulting MAP autocalibrating image reconstruction algorithm would be a direct analog of that developed using a velocity field prior in 4D MAP image reconstruction.

These efforts represent a general effort to improve our model of image acquisition. Currently, every 3D medical image is generated under a set of assumptions which, when broken, introduce alias artifacts that obfuscate important shape and motion information. Chapters 2 and 3 represent an effort to soften these assumptions and include their parameters in a model estimation framework. Thus there is clearly an opportunity to combine geometry calibration with internal anatomical motion estimation to provide full geometry and motion estimation during image formation.

Practically, autocalibrating CT image reconstruction is exciting because of its potential to enable 3D scans using existing widely deployed scanners. C-arm fluoroscopy is currently the main application of this technology, but it will be very interesting to push the limits, such as applying autocalibration to mobile x-ray. Such applications promise to bring 3D medical imaging to previously unthinkable applications, such as in ambulances, remote

rescue operations, and on the battlefield.

6.2.2 Polynomial Regression

Chapter 4 established that Riemannian polynomials are useful for curve regression, flexible enough to provide improved curve fitting over geodesics, while not introducing unnecessary complexity. Quadratic and cubic polynomials, in particular, were shown to outperform geodesics when applied to rat calivaria and human corpus callosum datasets. This suggests that Riemannian polynomials should be part of a standard toolbox for parametric curve regression on Riemannian manifolds, as their Euclidean counterparts are for conventional data analysis. However, accompanying statistical methods, in particular for model selection, will be necessary in order to evaluate polynomials in the context of geodesics and other models.

As in Euclidean space, model selection (the choice of polynomial order) is an immediate challenge when using Riemannian polynomials. In this dissertation, I have used R^2 to characterize goodness of fit, which gives information about underfitting but is insufficient for determining overfitting. Instead, other summary measures such as predicted R^2_{pred}, in which residuals are tested on polynomials fit to data sets in which they are excluded, may be preferable, but will necessarily require considerably more computation. On the other hand, information-based adjustment factors, such as the Akaike information criterion (AIC) or Bayes information criterion (BIC), may be able to be generalized to the Riemannian polynomial setting. Hypothesis testing for model selection is another possibility, but as with these other methods, it will require adaptation to be applicable on curved spaces.

The compact representation of polynomials will be useful when viewed in a statistical hypothesis driven framework. For instance, comparative analysis based on curve parameters promises to exploit the parametric nature of polynomials in a rigorous statistical setting. However, even though the data manifold may be well-understood, the space of parameters (a point in the manifold and a collection of tangent vectors at that point) may have considerably more complex geometry. For instance, Muralidharan and Fletcher [9] propose using Sasaki geometry for hypothesis testing of geodesic trends between populations in comparative longitudinal studies. The Sasaki metric is a natural way to endow the space of geodesic parameters (the tangent bundle) with Riemannian geometry necessary for statistical analysis, but computation of, for instance, geodesics is more complicated than on the original data space. It will be interesting to explore the generalization of this natural metric to *fibered products* of tangent bundles, which constitute the parameter spaces of Riemannian polynomials.

6.2.3 Irrotational Diffeomorphisms

Irrotational diffeomorphic shape analysis, presented in Chapter 5, is not the first example of medical image analysis making use of constrained subsets of diffeomorphisms. For instance, incompressible diffeomorphisms were shown in Chapter 2 to be of considerable use for motion modeling. However, the space $\mathrm{SDiff}(\Omega)$ of incompressible diffeomorphisms has been well-studied in the physical sciences [2]. Its geometry ($\mathrm{SDiff}(\Omega)$ is a well-formed Lie subgroup of $\mathrm{Diff}_c(\Omega)$), and many of its practical uses (it is the basis for the incompressible Euler equations in fluid dynamics) were known before quantitative medical image analysis was invented.

Discovery of $\mathrm{IDiff}(\Omega)$, is very exciting because it represents a new perspective on diffeomorphic shape analysis. The exploration of irrotational diffeomorphisms in a computational anatomy context is representative of a recent trend in which not only fluid dynamics but shape analysis also drives theoretical development of diffeomorphisms and differential geometry. The work in this dissertation makes no claim to providing those theoretical developments. Still, Chapter 5 serves not only as a development of considerable practical importance, but an important motivation for the study of Modin's polar factorisation and irrotational diffeomorphism in general.

Currently, efficient image registration and atlas-building algorithms have been developed on $\mathrm{IDiff}(\Omega)$, but there is clearly an opportunity for development of dynamic shape change algorithms such as geodesic and polynomial regression. As geodesics in $\mathrm{IDiff}(\Omega)$ are simply geodesics on the wider diffeomorphism group with a particular metric and with particular initial conditions, existing computational anatomy and regression algorithms can be used on $\mathrm{IDiff}(\Omega)$ as well, with little modification. This may be a useful first step toward developing geodesic and polynomial regression on $\mathrm{IDiff}(\Omega)$. It will be even more interesting, though, to exploit the flatness of $\mathrm{IDiff}(\Omega)$ to provide computationally efficient algorithms for regression, as extensions of image matching and atlas construction.

A major goal of irrotational diffeomorphic image analysis is to provide a new framework for longitudinal analysis. Current diffeomorphic longitudinal frameworks, such as that of Singh et al. [11], use the same diffeomorphic shape change model to describe both temporal evolution and intersubject variability. However, there is no reason to believe that in a longitudinal study both modes of deformation should be the same. Growth and disease progression are physical processes, while correspondence between subjects is not governed by biology but by a human assessment of similarity. Since the model for temporal change is the real object of interest, it is desirable to restrict growth and disease progression to

physically plausible modes of shape evolution. It is feasible that irrotational diffeomorphism is the predominant mode of shape evolution, since at least some aspects of growth are driven predominantly by volume change. This actually underlines the use of volume change in the analysis of deformation in previous anatomical studies [3, 1].

References

[1] P Aljabar et al. "Assessment of Brain Growth in Early Childhood Using Deformation-Based Morphometry". In: *Neuroimage* 39.1 (2008), pp. 348–358.

[2] Vladimir I Arnol'd. "Sur la Géométrie Différentielle des Groupes de Lie de Dimension Infinie et ses Applications à l'Hydrodynamique des Fluides Parfaits". In: *Ann. Inst. Fourier* 16 (1966), pp. 319–361.

[3] John Ashburner et al. "Identifying Global Anatomical Differences: Deformation-Based Morphometry". In: *Human Brain Mapping* 6.5-6 (1998), pp. 348–357.

[4] P. Thomas Fletcher. "Geodesic Regression on Riemannian Manifolds". In: *International Workshop on Mathematical Foundations of Computational Anatomy MFCA*. 2011.

[5] Sarah E Geneser et al. "Quantifying Variability in Radiation Dose Due to Respiratory-Induced Tumor Motion". In: *Medical Image Analysis* 15 (2011), pp. 640–649.

[6] Jacob Hinkle et al. "4D MAP MRI Image Reconstruction". In: *International Conference on Computer Vision Theory and Applications (VISAPP)*. Angers, France, May 2010.

[7] Peter E. Jupp and John T. Kent. "Fitting Smooth Paths to Spherical Data". In: *Appl. Statist.* 36.1 (1987), pp. 34–46.

[8] K. Modin. "Generalised Hunter-Saxton Equations, Optimal Information Transport, and Factorisation of Diffeomorphisms". In: *ArXiv e-prints* (Mar. 2012). arXiv: `1203.4463 [math-ph]`.

[9] Prasanna Muralidharan and P Thomas Fletcher. "Sasaki Metrics for Analysis of Longitudinal Data on Manifolds". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 1027–1034.

[10] Marc Niethammer, Yang Huang, and François-Xavier Vialard. "Geodesic Regression for Image Time-Series". In: *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2011.

[11] Nikhil Singh et al. "A Hierarchical Geodesic Model for Diffeomorphic Longitudinal Shape Analysis". In: *Information Processing in Medical Imaging (IPMI)*. Springer. 2013, pp. 560–571.

APPENDIX A

GEODESIC REGRESSION IN LIE GROUPS

In this appendix, I review the theory of geodesic regression in a Lie group with a right or left invariant metric. In particular, the Euler-Poincaré equation, which is the fundamental equation of geodesic evolution in a Lie group, and the reduced Jacobi equation, describing variations of geodesics, are derived. In each case, the resulting formulas will be shown to be quite simple, owing to the additional group structure in a Lie group compared to a general Riemannian manifold. Much of this appendix contains well-known results, derived in many geometric mechanics texts, for instance [10, 5, 9, 7, 3].

A.1 Lie Groups and Lie Algebras

In this appendix, G denotes a Lie group with identity $e \in G$ and Lie algebra $\mathfrak{g}$. This means that G is a group and also a smooth manifold, and that multiplication and inversion are smooth functions. It also means that $\mathfrak{g}$ is a vector space, which has the structure of an *associative algebra*, meaning that there is defined a skew-symmetric bilinear vector product ad, called the *infinitesimal adjoint action* of $\mathfrak{g}$ on itself:

$$\operatorname{ad} : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}. \tag{A.1}$$

In this context, the term *skew-symmetric* refers to the property that

$$\operatorname{ad}_v w = -\operatorname{ad}_w v, \tag{A.2}$$

for all vectors $v, w \in \mathfrak{g}$. That $\mathfrak{g}$ is a *Lie* algebra implies another condition called the Jacobi identity:

$$\operatorname{ad}_u \operatorname{ad}_v w + \operatorname{ad}_w \operatorname{ad}_u v + \operatorname{ad}_v \operatorname{ad}_w u = 0, \tag{A.3}$$

for all vectors $u, v, w \in \mathfrak{g}$.

A.1.1 Adjoint Representation

In this section, I will give a very brief derivation of the *adjoint representation* of a Lie group G over its Lie algebra $\mathfrak{g}$. The term *representation* refers to a smooth homomorphism from the group G into a group of linear transformations over the vector space $\mathfrak{g}$. This is analogous to the representation of finite groups as matrix groups, and matrix multiplication is a good guiding intuition to keep in mind while working with Lie groups. The derivation I present here is brief as I will try to impart the spirit of the derivation, but will omit a few details for clarity's sake. The reader is referred to any Lie group or geometric mechanics text for a more complete treatment of the material in this subsection, such as [10, 7, 1, 8].

Consider the group of *inner automorphisms* of a group G,

$$\mathrm{Inn}(G) = \{h \mapsto ghg^{-1} : g \in G\}. \tag{A.4}$$

Each inner automorphism corresponds to a process of *conjugation* by a group element $g \in G$. Notice that if the group G is Abelian, then $\mathrm{Inn}(G)$ consists of a single element, the identity, since commutativity makes every conjugation act trivially. Triviality of $\mathrm{Inn}(G)$ is actually also a *sufficient* condition for the group G to be Abelian. This hopefully gives some intuition that the inner automorphism group $\mathrm{Inn}(G)$ gives information about the *degree* to which the group G is Abelian.

I will denote by Φ_h the inner automorphism corresponding to group element $h \in G$:

$$\Phi_g(h) = ghg^{-1}. \tag{A.5}$$

Consider a curve $h(t)$, which is the identity at time zero and whose derivative at that time is a vector $\eta \in \mathfrak{g}$. Since $\Phi_g(e) = e$, the derivative of $\Phi_g(h(t))$ is an element of the Lie algebra as well. This derivative is what is called the *adjoint action* of the element g on the Lie algebra element η, and is written

$$\mathrm{Ad}_g\,\eta = \frac{d}{dt}\Big|_{t=0}\Phi_g(h(t)). \tag{A.6}$$

In terms of right and left translation, the adjoint action of g on η is computed by left translating η by g, then right translating the result back to the identity using g^{-1}, or by doing these same operations in the reverse order. Thus I will often use the shorthand

$$\mathrm{Ad}_g\,\eta = g\eta g^{-1}. \tag{A.7}$$

The group action $\mathrm{Ad} : G \times \mathfrak{g} \to \mathfrak{g}$ is fundamental to the study of the Lie group G. For every $g \in G$, Ad_g is a linear mapping $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$, and so the functor Ad maps Lie groups into linear operators on vector spaces. That mapping is differentiable once more.

Consider now a curve $g(s) \in G$, again with $g(0) = e$, and with derivative at time zero $\xi \in \mathfrak{g}$. Since $\mathrm{Ad}_{g(s)}\,\eta \in \mathfrak{g}$ and $\mathfrak{g}$ is a vector space, its derivative is also in $\mathfrak{g}$. We call this derivative the *infinitesimal adjoint action* and write

$$\mathrm{ad}_\xi\,\eta = \frac{d}{ds}\Big|_{s=0} \mathrm{Ad}_{g(s)}\,\eta. \tag{A.8}$$

Using the shorthand above for $\mathrm{Ad}_g\,\eta$ and the intuition from matrix multiplication, we can write the following:

$$\mathrm{ad}_\xi\,\eta = \frac{d}{ds}\Big|_{s=0} g(s)\eta g(s)^{-1} \tag{A.9}$$

$$= \xi\eta - \eta\xi. \tag{A.10}$$

Note that this is very similar to the *matrix commutator*, and as such it also is clearly skew-symmetric in ξ and η. Also, the infinitesimal adjoint action, $\mathrm{ad} : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, gives a sense of the degree to which the Lie group G is Abelian. This is clear since, as noted previously if G is Abelian, $\mathrm{Inn}(G)$ is trivial, so all curves considered in this argument are constant and all derivatives are zero.

A.1.2 Dual Adjoint Representation

As $\mathfrak{g}$ is a vector space, it is accompanied by its dual space $\mathfrak{g}^*$, which consists of all linear functionals $\mathfrak{g} \to \mathbb{R}$. We denote the pairing of a linear functional on a vector via round braces:

$$(\mu, v) = \mu(v) \in \mathbb{R} \qquad \forall \mu \in \mathfrak{g}^*, \quad v \in \mathfrak{g}. \tag{A.11}$$

Using this dual pairing, the dual of any linear function $f : \mathfrak{g} \to \mathfrak{g}$ is defined as the function $f^* : \mathfrak{g}^* \to \mathfrak{g}^*$ satisfying, for all $v \in \mathfrak{g}, \mu \in \mathfrak{g}^*$,

$$(f^*\mu, v) = (\mu, fv). \tag{A.12}$$

Thus, the *coadjoint action* $\mathrm{Ad}^*_g : \mathfrak{g}^* \to \mathfrak{g}^*$ is the dual of $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$ and the *infinitesimal coadjoint action*, $\mathrm{ad}^*_v : \mathfrak{g}^* \to \mathfrak{g}^*$ is the dual of $\mathrm{ad}_v : \mathfrak{g} \to \mathfrak{g}$.

A.2 Invariance

As G is a smooth manifold, for every point $g \in G$, there exists a vector space T_gG which is the tangent space to G at g. Any tangent vector $v \in T_gG$, by definition, is the derivative of a curve $\gamma : (-\epsilon, \epsilon) \to G$ satisfying $\gamma(0) = g$. Now fix some element $h \in G$ and consider the curve $h\gamma(t)$. This is a curve passing through $hg \in G$ at time $t = 0$. Thus the derivative

is a vector in $T_{hg}G$, referred to as the *left translate* of v by h, denoted hv. Right translation by h of v, vh, is defined similarly, by considering the derivative of the curve $\gamma(t)h$,.

Now imagine we are given a vector field $X : G \to TG$. For every point $g \in G$, define a vector in T_eG by left translation of $X(g)$ by the element g^{-1} (notice that this is necessarily a tangent vector at $g^{-1}g = e$). If, for all g, these vectors $g^{-1}X(g)$ are equal to the *same* element in T_eG, then the vector field X is said to be *left invariant*.[1]

As G is a smooth manifold, there exist many choices of Riemannian metric one might place on G. These are determined by a choice of smoothly spatially-varying inner product on the tangent spaces T_gG. On Lie groups, certain classes of metrics offer advantages since they exploit the group structure and in particular the ability to *translate* vectors, which is not possible on a general manifold. These are the left and right invariant metrics. For example, a left invariant metric is one such that for any tangent vectors $v, w \in T_gG$ and for any group element h,[2]

$$\langle v, w \rangle_g = \langle hv, hw \rangle_{hg}. \tag{A.13}$$

The extremely useful property of left and right invariant metrics is that their definition is determined entirely by the inner product at *any element* $g \in G$, since any other inner product is computed by translation to that special point. The tangent space at the identity $T_eG \cong \mathfrak{g}$, is a particularly convenient choice of point at which to define an invariant metric. In this case, the metric is defined as a symmetric bilinear positive-definite mapping $\mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$. This is equivalent to a choice of so-called *inertia operator* $L : \mathfrak{g} \to \mathfrak{g}^*$ using the definition

$$\langle v, w \rangle = (Lv, w) \qquad \forall v, w \in \mathfrak{g}. \tag{A.14}$$

Note that in order to satisfy the properties of a well-formed Riemannian metric, L must be invertible and

$$(Lv, w) = (Lw, v) \qquad \forall v, w \in \mathfrak{g}. \tag{A.15}$$

Equipped with an inner product on $\mathfrak{g}$, a linear operator $f : \mathfrak{g} \to \mathfrak{g}$ is *transposed*[3] with respect to the inner product defined by L, using the formula

---

[1] As will be common in this section, an analogous definition exists for right invariance. When these definitions do not offer any surprises, they will not be mentioned, for simplicity's sake.

[2] Note that I subscript the inner products here to emphasize that these are realizations of the Riemannian metric at different points $g, hg \in G$. I will not usually include these subscripts, for simplicity.

[3] The definition given here is that of the adjoint associated to a linear operator with respect to an inner product. However, we refer to it as a transpose, reminiscent of the transpose of a matrix, due to the unfortunate conflict of terminology with the adjoint representation.

$$\langle f^\dagger v, w\rangle = \langle v, fw\rangle \qquad \forall v, w \in \mathfrak{g}. \tag{A.16}$$

This we use to define the adjoint-transpose action $\mathrm{Ad}^\dagger : G \times \mathfrak{g} \to \mathfrak{g}$ via the transpose of Ad_g and the infinitesimal adjoint-transpose $\mathrm{ad}^\dagger : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ via the transpose of ad_v.

A.3 Covariant Derivatives

In this section, we fix a particular left invariant metric, determined by left translation and an inertia operator L, and derive a convenient formula for the Levi-Civita connection using that metric. First, we introduce two operators that are rarely seen (c.f. Lewis and Murray [9] and Bullo [4]) but greatly simplifies computations in the rest of this appendix. The first operator is called the *symmetric product*, $\mathrm{sym} : \mathfrak{g} \to \mathfrak{g} \to \mathfrak{g}$:

$$\mathrm{sym}_v\, w := -\left(\mathrm{ad}^\dagger_v\, w + \mathrm{ad}^\dagger_w\, v\right). \tag{A.17}$$

Using this, we define

$$\overline{\nabla}_v w := \frac{1}{2}\left(\mathrm{ad}_v\, w + \mathrm{sym}_v\, w\right). \tag{A.18}$$

Given left invariant vector fields corresponding to $v, w \in \mathfrak{g}$, the covariant derivative of gv along gw is a left invariant vector field whose left trivialization is

$$g^{-1}\nabla_{gv}(gw) = \overline{\nabla}_v w. \tag{A.19}$$

For a proof of this, see for instance Bullo [4, Thm. 2]. Notice that the operator $\overline{\nabla}$ in this case can be interpreted as a *reduced* Levi-Civita connection. Care must be taken with this interpretation, as a negative sign is obtained in the case of right invariance.

A.3.1 Geodesic Equation

Note that using the geodesic equation from Riemannian geometry [5] and these formulas for covariant derivatives, the reduced geodesic equation could easily be derived by substitution. In order to give more intuition for the variational derivation using Lie group methods, in the following section, a general version of these equations will be derived.

A.4 Euler-Poincaré Reduction

Suppose we are given a Lagrangian, meaning a differentiable function $\mathcal{L} : TG \to \mathbb{R}$. $\mathcal{L}$ is called left invariant if, for all $h \in G, (g, v) \in TG$:

$$\mathcal{L}(hg, hv) = \mathcal{L}(g, v). \tag{A.20}$$

For instance, a left invariant metric on G gives rise to a left invariant Lagrangian:

$$\mathcal{L}(g, v) = \frac{1}{2}\|v\|_g^2, \tag{A.21}$$

which could represent the kinetic energy of a physical system, for example.

A choice of left invariant Lagrangian is equivalent to a choice of Lagrangian on the Lie algebra, which we denote $\ell : \mathfrak{g} \to \mathbb{R}$:

$$\mathcal{L}(e, X) = \mathcal{L}(g, gX) =: \ell(X), \quad \forall g \in G, X \in \mathfrak{g}. \tag{A.22}$$

Now let g depend on time t and define the action functional, S:

$$S[g] = \int_0^1 \mathcal{L}(g(t), \dot{g}(t))dt. \tag{A.23}$$

For a Lagrangian like the one in (A.21), minimization of S, fixing the endpoints, is equivalent to computation of a geodesic connecting those points [5, 7].

I will use the notation $\xi \in \mathfrak{g}$ for the left trivialized velocity of g:

$$\xi(t) = g^{-1}(t)\dot{g}(t). \tag{A.24}$$

Because of left invariance, the action functional takes the form

$$S = \int_0^1 \ell(\xi(t))dt. \tag{A.25}$$

To derive the Euler-Lagrange equations, we proceed in the usual way on a smooth manifold.

Let $\delta g(t) \in T_g G$ be a variation in g which is zero at $0, 1$. This means we choose a smooth family of curves $\{g_s\}, s \in (-\epsilon, \epsilon)$ for which the following conditions all hold

$$g_0 = g \tag{A.26}$$

$$\frac{d}{ds}|_{s=0} g_s(t) = \delta g(t). \tag{A.27}$$

If $(\delta S, \delta g)$ in (A.23) is zero for all such variations, the curve g is extremal in $\mathcal{L}$ among curves fixing $g(0), g(1)$ and so could, for instance, be a geodesic depending on the form of $\mathcal{L}$. This is equivalent to the variation of (A.25) being zero for variations $\delta\xi : [0, 1] \to \mathfrak{g}$ of the form

$$\delta\xi = \dot{Z} + \mathrm{ad}_\xi Z, \tag{A.28}$$

for all function $Z : \mathbb{R} \to \mathfrak{g}$ with $Z(0) = Z(1) = 0$. To see this, define Z to be the left trivialization of δg,

$$Z = g^{-1}\delta g. \tag{A.29}$$

The variation of ξ and the time derivative of Z are derived using the product rule:[4]

$$\delta\xi = \delta(g^{-1}\dot{g}) = g^{-1}\dot{\delta g} - g^{-1}\delta g.g^{-1}\dot{g} \tag{A.30}$$

$$\dot{Z} = \frac{d}{dt}(g^{-1}\delta g) = g^{-1}\dot{\delta g} - g^{-1}\dot{g}.g^{-1}\delta g \tag{A.31}$$

$$\delta\xi = \dot{Z} + g^{-1}\dot{g}.g^{-1}\delta g - g^{-1}\delta g.g^{-1}\dot{g} \tag{A.32}$$

$$= \dot{Z} + \xi.Z - Z.\xi = \dot{Z} + \mathrm{ad}_\xi\, Z, \tag{A.33}$$

proving the assertion that the right-hand side above is the corresponding variation $\delta\xi$, corresponding to δg.

Finally, take the variation of S with respect to a variation of this form:

$$\int_0^1 \left(\frac{\delta S}{\delta g}, \delta g\right) dt = \int_0^1 \left(g^{-1}\frac{\delta S}{\delta g}, Z\right) dt \tag{A.34}$$

$$= \int_0^1 \left(\frac{\delta \ell}{\delta \xi}, \delta\xi\right) dt \tag{A.35}$$

$$= \int_0^1 \left(\frac{\delta \ell}{\delta \xi}, \dot{Z} + \mathrm{ad}_\xi\, Z\right) dt \tag{A.36}$$

$$= \int_0^1 \left(-\frac{d}{dt}\frac{\delta \ell}{\delta \xi} + \mathrm{ad}^*_\xi \frac{\delta \ell}{\delta \xi}, Z\right) dt \tag{A.37}$$

where we integrated by parts using the fact that Z vanishes at $0, 1$. Since this holds for all choices of variations Z, we have the Euler-Lagrange equation

$$\frac{d}{dt}\frac{\delta \ell}{\delta \xi} = \mathrm{ad}^*_\xi \frac{\delta \ell}{\delta \xi}. \tag{A.38}$$

This equation is commonly called the left invariant Euler-Poincaré equation

Note that I am talking about momenta and covectors and relating it to an evolution equation for a curve, without ever explicitly talking about a Riemannian metric. Note that the Euler-Poincaré applies to any left invariant Lagrangian. For a purely kinetic Lagrangian,

$$\ell(\xi) = \frac{1}{2}\|\xi\|^2 = \frac{1}{2}(L\xi, \xi) \tag{A.39}$$

where L is the inertia operator corresponding to the metric, the partial derivative with respect to ξ takes a particularly familiar form:

$$\frac{\delta \ell}{\delta \xi} = L\xi. \tag{A.40}$$

If L is truly an inertia operator, or metric, on $\mathfrak{g}$, meaning it is invertible, then we can recover the evolution equation for ξ from the Euler-Poincaré equation:

$$\frac{d}{dt}\xi = L^{-1}\,\mathrm{ad}^*_\xi(L\xi) = \mathrm{ad}^\dagger_\xi\,\xi. \tag{A.41}$$

[4]Note that a rigorous derivation of this requires the use of differentials (pushforwards) of the left and right translation maps. This considerably clutters the notation and offers no surprises, so for clarity's sake, in this derivation, I use simple concatenation notation and familiar product rules from matrix multiplication.

A.4.1 Right Invariant Euler-Poincaré Equation

Note that in the above section, only left invariance was considered. In order to derive the right invariant Euler-Poincaré equation, the same general argument is used, but instead of left invariance and left trivialization, their counterparts are substituted. Defining the velocity as

$$\xi(t) = \dot{g}(t)g^{-1}(t) \tag{A.42}$$

and

$$Z(t) = \delta g(t)g^{-1}(t), \tag{A.43}$$

we are led to the alternative formula for $\delta\xi$:

$$\delta\xi = \dot{Z} - \mathrm{ad}_{\xi} Z. \tag{A.44}$$

Note that this formula differs from (A.28) only in the negative sign. To see where the negative sign comes from, we derive the formula just as before but using right translation:

$$\delta\xi = \delta(\dot{g}g^{-1}) = \dot{\delta g}g^{-1} - \dot{g}g^{-1}\delta g.g^{-1} \tag{A.45}$$

$$\dot{Z} = \frac{d}{dt}(\delta g g^{-1}) = \dot{\delta g}g^{-1} - \delta g g^{-1}\dot{g}.g^{-1} \tag{A.46}$$

$$\delta\xi = \dot{Z} + \delta g g^{-1}\dot{g}.g^{-1} - \dot{g}g^{-1}\delta g.g^{-1} \tag{A.47}$$

$$= \dot{Z} + Z.\xi - \xi.Z = \dot{Z} - \mathrm{ad}_{\xi} Z. \tag{A.48}$$

This leads immediately to the right invariant version of (A.37):

$$\int_0^1 \left(\frac{\delta S}{\delta g}, \delta g\right) dt = \int_0^1 \left(-\frac{d}{dt}\frac{\delta \ell}{\delta \xi} - \mathrm{ad}^*_{\xi}\frac{\delta \ell}{\delta \xi}, Z\right) dt \tag{A.49}$$

which in turn provides the *right invariant* Euler-Poincaré equation:

$$\frac{d}{dt}\frac{\delta \ell}{\delta \xi} = -\mathrm{ad}^*_{\xi}\frac{\delta \ell}{\delta \xi}. \tag{A.50}$$

A.4.2 Integrated Form of Euler-Poincaré Equation

As discussed, the following two equations describe geodesics in a Lie group with left invariant metric:

$$\frac{d}{dt}g(t) = g(t)\xi(t) \qquad \in T_{g(t)}G \tag{A.51}$$

$$\frac{d}{dt}\xi(t) = \mathrm{ad}^{\dagger}_{\xi}\xi(t) \qquad \in \mathfrak{g}. \tag{A.52}$$

The first equation is the definition of ξ as the left trivialized velocity, while the second equation gives the evolution of ξ, the left invariant Euler-Poincaré equation. In practice,

numerical integration of the first equation depends on aspects of G [2]. Since ξ resides in a vector space, conventional methods could be used to integrate the Euler-Poincaré equation. However, propagating errors in ξ will necessarily compound errors in $g(t)$. In this section, I will derive a useful formula for exact integration of ξ, given $g(t)$, which will provide a sound basis for practical integration of geodesics in Lie groups.

Letting $\eta \in \mathfrak{g}$ be some fixed vector, I will prove the following:

$$\frac{d}{dt}\operatorname{Ad}^*_{g(t)}\eta = \operatorname{ad}^*_{g^{-1}\dot{g}}\operatorname{Ad}^*_g\eta \tag{A.53}$$

$$\frac{d}{dt}\operatorname{Ad}^*_{g(t)^{-1}}\eta = -\operatorname{ad}^*_{\dot{g}g^{-1}}\operatorname{Ad}^*_{g^{-1}}\eta. \tag{A.54}$$

Again I will use the imprecise, but compact notation corresponding to matrix Lie groups:

$$\operatorname{Ad}_g v = gvg^{-1} \tag{A.55}$$

$$\operatorname{ad}_u v = uv - vu \tag{A.56}$$

$$\operatorname{Ad}^*_g\eta = g^{-1}\eta g \tag{A.57}$$

$$\operatorname{ad}^*_u\eta = -u\eta + \eta u. \tag{A.58}$$

Using the product rule and inverse derivative rule from multivariate calculus, we derive

$$\frac{d}{dt}\operatorname{Ad}^*_{g(t)}\eta = -g^{-1}\dot{g}g^{-1}\eta g + g^{-1}\eta\dot{g} \tag{A.59}$$

$$= -(g^{-1}\dot{g})g^{-1}\eta g + g^{-1}\eta g g^{-1}\dot{g} \tag{A.60}$$

$$= \operatorname{ad}^*_{g^{-1}\dot{g}}\operatorname{Ad}^*_g\eta \tag{A.61}$$

$$\frac{d}{dt}\operatorname{Ad}^*_{g^{-1}}\eta = \frac{d}{dt}\left(g\eta g^{-1}\right) \tag{A.62}$$

$$= \dot{g}\eta g^{-1} - g\eta g^{-1}\dot{g}g^{-1} \tag{A.63}$$

$$= (\dot{g}g^1)g\eta g^{-1} - g\eta g^{-1}(\dot{g}g^{-1}) \tag{A.64}$$

$$= -\operatorname{ad}^*_{\dot{g}g^{-1}}\operatorname{Ad}^*_{g^{-1}}\eta \tag{A.65}$$

This proves the claim above.

The first relation above, in the left invariant case, satisfies the equation

$$\frac{d}{dt}\mu = \operatorname{ad}^*_\xi\mu \tag{A.66}$$

if we use the definition

$$\mu(t) = \operatorname{Ad}^*_{g(t)}\eta. \tag{A.67}$$

Notice that when $\eta = L\xi(0)$ is the initial *momentum* of a geodesic, it is easily verified that $L\xi = \mu$ satisfies

$$\frac{d}{dt}\xi = \operatorname{ad}^\dagger_\xi\xi. \tag{A.68}$$

This implies that the velocity of a geodesic in the left invariant case is given by

$$\xi(t) = \mathrm{Ad}^{\dagger}_{g(t)}\,\xi(0), \tag{A.69}$$

which can be considered the *integrated form* of the left invariant Euler-Poincaré equation. The right invariant formula is derived similarly and is given by

$$\xi(t) = \mathrm{Ad}^{\dagger}_{g(t)^{-1}}\,\xi(0). \tag{A.70}$$

A.5 The Reduced Jacobi Equation

The Jacobi equation describes geodesic variations on a Riemannian manifold M (c.f. [5, Chap. 5] and [6]). That is, consider being handed a differentiable family of geodesics parametrized by some parameter s, so that for every s, the curve $\gamma_s(t)$ is a geodesic. The derivative at $s = 0$ of that family of curves provides a vector field along the curve $\gamma_0(t)$, which is a *geodesic variation* of γ_0:

$$\delta\gamma(t) \in T_{\gamma(t)}M. \tag{A.71}$$

Such a vector field is called a *Jacobi field*. Just as any geodesic must satisfy a particular ordinary differential equation (the geodesic or Euler-Poincaré equation), so must any Jacobi field. The name of that equation, naturally, is the *Jacobi equation*, and on a general Riemannian manifold, it is given by

$$\nabla_{\dot\gamma}\nabla_{\dot\gamma}\delta\gamma = R(\delta\gamma, \dot\gamma)\dot\gamma, \tag{A.72}$$

where R denotes the Riemann curvature tensor associated with the manifold M. For a derivation of the Jacobi equation and its generalization in the case of polynomials, refer to Appendix 4.10 of Chapter 4 on page 55.

The Jacobi equation could be adapted directly to Lie groups by introducing left trivialized vectors. Doing so directly, one would introduce variables to represent the left trivialized Jacobi field $Z = \gamma^{-1}\dot\gamma$ and acceleration $\eta = \gamma^{-1}\nabla_{\dot\gamma}\delta\gamma$. This leads to considerably complicated formulas when inserted into the Jacobi equation directly. Instead, by taking a variation directly, in the presence of the Euler-Poincaré equation, we will derive a much simpler *reduced Jacobi equation*. Note that this is not a well-known method and the representation below for reduced Jacobi fields is not widely used, but it is my opinion that it should be. I was first exposed to it in the excellent report Bullo [4], which is related to the publication Lewis and Murray [9], both of which the reader is strongly recommended to investigate.

We still define the vector $Z(t) \in \mathfrak{g}$ as a left trivialized perturbation of γ:

$$Z(t) = \gamma(t)^{-1}\delta\gamma(t). \tag{A.73}$$

Recall that this induces a variation in the left trivialized velocity ξ:

$$\delta\xi = \frac{\partial}{\partial t}Z + \operatorname{ad}_\xi Z. \tag{A.74}$$

Also recall that for geodesics, the velocity must satisfy

$$\frac{\partial}{\partial t}\xi = \operatorname{ad}^\dagger_\xi \xi. \tag{A.75}$$

The right-hand side of this equation is bilinear in ξ. Remember that the variation δ is a (partial) differential operator, corresponding to the derivative within a geodesic family, so it satisfies the product rule. This lets us write[5]

$$\frac{\partial}{\partial t}\delta\xi = \operatorname{ad}^\dagger_{\delta\xi}\xi + \operatorname{ad}^\dagger_\xi \delta\xi = -\operatorname{sym}_\xi \delta\xi. \tag{A.76}$$

To summarize, the perturbation Z representing a left invariant reduced Jacobi field satisfies the following system of equations:

$$\frac{\partial}{\partial t}Z = \delta\xi - \operatorname{ad}_\xi Z \tag{A.77}$$

$$\frac{\partial}{\partial t}\delta\xi = \operatorname{ad}^\dagger_{\delta\xi}\xi + \operatorname{ad}^\dagger_\xi \delta\xi = -\operatorname{sym}_\xi \delta\xi. \tag{A.78}$$

This is best written in block matrix form:

$$\frac{\partial}{\partial t}\begin{pmatrix} Z \\ \delta\xi \end{pmatrix} = \begin{pmatrix} -\operatorname{ad}_\xi & I \\ 0 & -\operatorname{sym}_\xi \end{pmatrix}\begin{pmatrix} Z \\ \delta\xi \end{pmatrix} \tag{A.79}$$

This is an incredibly simple formula for computing the Jacobi equation on a Lie group with left invariant metric.

Note that using the right invariant version of the Euler-Poincaré equation and right trivialization, the following similar formula is obtained for right invariant Jacobi fields:

$$\frac{\partial}{\partial t}\begin{pmatrix} Z \\ \delta\xi \end{pmatrix} = \begin{pmatrix} \operatorname{ad}_\xi & I \\ 0 & \operatorname{sym}_\xi \end{pmatrix}\begin{pmatrix} Z \\ \delta\xi \end{pmatrix}. \tag{A.80}$$

Note that in both this case and the left invariant one, the variables Z and $\delta\xi$ satisfy a time-dependent linear ODE of the form:

$$\frac{\partial}{\partial t}\begin{pmatrix} Z \\ \delta\xi \end{pmatrix} = A(t)\begin{pmatrix} Z \\ \delta\xi \end{pmatrix} \tag{A.81}$$

for some operator $A(t)$ written in block matrix form. As mentioned in Appendix 4.10 of Chapter 4 on page 55, *adjoint Jacobi fields*, which are used to propagate adjoint variables

[5]The left-hand side in this equation makes use of the equality of mixed partials: $\frac{\partial}{\partial t}\frac{\partial}{\partial s} = \frac{\partial}{\partial s}\frac{\partial}{\partial t}$.

necessary for geodesic regression, are obtained by simply computing the adjoint of this ODE, resulting in an ODE with transition matrix $-A^\dagger(t)$.

Table A.1 summarizes all the formulas related to geodesics in Lie groups with left and right invariant metrics.

A.6 Geodesic Regression and Adjoint Jacobi Fields

For geodesic regression, the useful equations are Euler-Poincaré and the *adjoint* Jacobi equation [11, 6]. In this section, I will show that surprisingly, these two equations are closely related. First, recall the left invariant Euler-Poincaré equation:

$$\dot{\xi} = \operatorname{ad}_\xi^\dagger \xi. \tag{A.82}$$

The matrix form ODEs were convenient for computing adjoint equations, but now we "unwrap" them into the following system. Introducing adjoint variables $\lambda, \mu \in \mathfrak{g}$, we read off the adjoint Jacobi equations from Table A.1:

$$\dot{\lambda} = \operatorname{ad}_\xi^\dagger \lambda \tag{A.83}$$

$$\dot{\mu} = -\lambda + \operatorname{sym}_\xi^\dagger \mu. \tag{A.84}$$

A specific formula for $\operatorname{sym}_\xi^\dagger$ is derived in Appendix 4.10 of Chapter 4.

The interesting part of this system is the first equation; it appears that λ follows a very similar evolution equation to that of ξ. Indeed, if $g(0) = e$, the adjoint variable μ satisfies

$$\mu(t) = \operatorname{Ad}_{g(t)}^\dagger \mu(0), \tag{A.85}$$

which is very similar to the integrated form of the Euler-Poincaré equation, (A.69). This gives a simple method for integration of the adjoint Jacobi equations necessary for geodesic regression [6]. These formulas are summarized in Table A.2.

Table A.1. The Rosetta stone of left and right invariant Lie group geodesic formulas.

	Left Invariant	Right Invariant
Reduced Levi-Civita	$\overline{\nabla}$	$-\overline{\nabla}$
Trivialized Velocity	$\xi := g^{-1}\dot{g}$	$\xi := \dot{g}g^{-1}$
Trivialized Jacobi Field	$Z := g^{-1}\delta g$	$Z := \delta g g^{-1}$
Induced Variation	$\delta\xi = \dot{Z} + \operatorname{ad}_\xi Z$	$\delta\xi = \dot{Z} - \operatorname{ad}_\xi Z$
Euler-Poincaré	$\dot{\xi} = \operatorname{ad}_\xi^\dagger \xi$	$\dot{\xi} = -\operatorname{ad}_\xi^\dagger \xi$
Jacobi Equation $A(t)$	$\begin{pmatrix} -\operatorname{ad}_\xi & I \\ 0 & -\operatorname{sym}_\xi \end{pmatrix}$	$\begin{pmatrix} \operatorname{ad}_\xi & I \\ 0 & \operatorname{sym}_\xi \end{pmatrix}$
Adjoint $-A^\dagger(t)$	$\begin{pmatrix} \operatorname{ad}_\xi^\dagger & 0 \\ -I & \operatorname{sym}_\xi^\dagger \end{pmatrix}$	$\begin{pmatrix} -\operatorname{ad}_\xi^\dagger & 0 \\ -I & -\operatorname{sym}_\xi^\dagger \end{pmatrix}$

Table A.2. Integration formulas for velocity and adjoint variables.

	Left Invariant	Right Invariant
Initial Velocity	$\xi_e = \mathrm{Ad}_{g^{-1}(0)}\,\xi(0)$	$\xi_e = \mathrm{Ad}_{g(0)}\,\xi(0)$
Velocity	$\xi(t) = \mathrm{Ad}^\dagger_{g(t)}\,\xi_e$	$\xi(t) = \mathrm{Ad}^\dagger_{g^{-1}(t)}\,\xi_e$
"Initial" λ	$\lambda_i = \mathrm{Ad}^\dagger_{g^{-1}(1)}\,\lambda(1)$	$\lambda_i = \mathrm{Ad}^\dagger_{g(1)}\,\lambda(1)$
Adjoint Jacobi Equations	$\lambda(t) = \sum_{t_i>t} \mathrm{Ad}^\dagger_{g(t)}\,\lambda_i$	$\lambda(t) = \sum_{t_i>t} \mathrm{Ad}^\dagger_{g^{-1}(t)}\,\lambda_e$
	$\dot\mu(t) = \mathrm{sym}^\dagger_{\xi(t)}\,\mu(t) - \lambda(t)$	$\dot\mu(t) = -\mathrm{sym}^\dagger_{\xi(t)}\,\mu(t) - \lambda(t)$

A.7 Matrix Lie Groups

A fundamental collection of Lie groups are the finite-dimensional matrix Lie groups, whose elements are square matrices. In this section, I will translate the results from the previous sections into the language of matrices and give a detailed account of the three-dimensional rotation group SO(3).

Let $\mathcal{M}(n)$ denote the monoid consisting of $n \times n$ real matrices under matrix multiplication. The simplest subgroup of $\mathcal{M}(n)$ is simply the collection of those matrices which are invertible (which is equivalent to having nonzero determinant):

$$\mathrm{GL}(n) = \{A \in \mathcal{M}(n) : \det A \neq 0\}. \tag{A.86}$$

This group is called the general linear group of degree n. It is a Lie group, locally smoothly homeomorphic to $\mathbb{R}^{n^2}$.

Common Lie subgroups are formed by imposing additional restrictions on $\mathrm{GL}(n)$. For instance, the orthogonal group $\mathrm{O}(n)$ consists of those elements A of $\mathrm{GL}(n)$ for which $A^T A = I$. Additionally, the special linear group, $\mathrm{SL}(n)$, consists of elements of $\mathrm{GL}(n)$ whose determinant is one.

Of the matrix Lie groups, SO(3) is of particular interest in shape analysis as it represents rotations of objects in $\mathbb{R}^3$. In the following sections, the necessary ingredients for computing dynamics on general matrix Lie groups will be derived, followed by a much more detailed investigation of the rotation group SO(3).

A.7.1 Matrix Lie Group Adjoint Representation

Suppose $G \leq \mathrm{GL}(n)$ is a general finite dimensional matrix Lie group. Elements of G are $n \times n$ matrices and multiplication is denoted by concatenation, so that an inner automorphism in G has the form

$$\Phi_A(B) = ABA^{-1}. \tag{A.87}$$

Clearly, this expression is linear in B, so the derivative at the identity is simply

$$\mathrm{Ad}_A\, Y = AYA^{-1}, \tag{A.88}$$

where C is an element of the Lie algebra $\mathfrak{g}$. Note that since G consists of square matrices, the set $\mathfrak{g}$ of all derivatives of such matrices are themselves $n \times n$ matrices, though the dimension of $\mathfrak{g}$ may be much smaller than n^2.

To derive the infinitesimal adjoint action, $\mathrm{ad}_X\, Y$, let $A(t) \in G$ be a differentiable family of matrices in G with

$$A(0) = I \tag{A.89}$$

$$\frac{d}{dt}|_{t=0} A(t) = X \tag{A.90}$$

and take the derivative of $\mathrm{Ad}_{A(t)}\, X$ at $t = 0$. Recall the following formula for the derivative of a matrix inverse:

$$\frac{d}{dt} A(t)^{-1} = -A(t)^{-1} \left(\frac{d}{dt} A(t) \right) A(t)^{-1}. \tag{A.91}$$

Applying this, along with the product rule, yields

$$\mathrm{ad}_X\, Y = \frac{d}{dt}|_{t=0}\, \mathrm{Ad}_{A(t)}\, Y \tag{A.92}$$

$$= \frac{d}{dt}|_{t=0} (A(t) Y A(t)^{-1}) \tag{A.93}$$

$$= XY - YX. \tag{A.94}$$

Note that this is clearly skew-symmetric. This quantity is commonly called the matrix commutator, $[X, Y]$. As a result of this general formula for $\mathrm{ad}_X\, Y$ in a matrix group, the infinitesimal adjoint action is commonly called the Lie bracket, and is sometimes written $[X, Y]$ for general Lie groups. This is confusing, however, when discussing vector fields, for which the Jacobi-Lie bracket is also defined. For that reason, in this work, the infinitesimal adjoint action is always referred to as $\mathrm{ad}_X\, Y$.

To derive the conjugate actions Ad^* and ad^*, we first identify the Lie coalgebra, $\mathfrak{g}^*$, consisting of all linear functionals over $n \times n$ real matrices. Elements of $\mathfrak{g}^*$ are also written

as $n \times n$ real matrices, with the dual pairing between matrices given by the usual Frobenius inner product of matrices. That is, for $X \in \mathfrak{g}$ and $\mu \in \mathfrak{g}^*$, we define

$$(\mu, X) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mu_i^j X_i^j = \mathrm{tr}(\mu^T X). \tag{A.95}$$

To compute the dual of Ad_A, let $\mu \in \mathfrak{g}^*$ and $Y \in \mathfrak{g}$, then conjugate with respect to the dual pairing between $\mathfrak{g}^*$ and $\mathfrak{g}$.

$$\begin{aligned}
(\mu, \mathrm{Ad}_A Y) &= (\mathrm{Ad}_A^* \mu, Y) && \text{(A.96)}\\
&= \mathrm{tr}(\mu^T A Y A^{-1}) && \text{(A.97)}\\
&= \mathrm{tr}(A^{-1} \mu^T A Y) && \text{(A.98)}\\
&= \mathrm{tr}((A^T \mu A^{-T})^T Y) && \text{(A.99)}\\
\mathrm{Ad}_A^* \mu &= A^T \mu A^{-T} && \text{(A.100)}
\end{aligned}$$

where A^{-T} indicates the inverse transpose of the matrix A. We have also used the fact that the trace of a product of matrices is invariant under cyclic permutation of the factors. The infinitesimal adjoint action is derived similarly:

$$\begin{aligned}
(\mu, \mathrm{ad}_X Y) &= (\mathrm{ad}_X^* \mu, Y) && \text{(A.101)}\\
&= \mathrm{tr}(\mu^T (XY - YX)) && \text{(A.102)}\\
&= \mathrm{tr}(\mu^T XY) - \mathrm{tr}(\mu^T YX) && \text{(A.103)}\\
&= \mathrm{tr}((X^T \mu)^T Y) - \mathrm{tr}(X \mu^T Y) && \text{(A.104)}\\
&= \mathrm{tr}((X^T \mu - \mu X^T)^T Y) && \text{(A.105)}\\
\mathrm{ad}_X^* \mu &= X^T \mu - \mu X^T. && \text{(A.106)}
\end{aligned}$$

A.7.2 The Rotation Group

The rotation group, SO(3), is presented as

$$\mathrm{SO}(3) = \{R \in \mathcal{M}(n) : R^T R = I, |R| = 1\}. \tag{A.107}$$

In order to derive the form of elements in its Lie algebra, $\mathfrak{so}(3)$, take a generic curve $R(t)$ through the identity in SO(3) with derivative $X \in \mathfrak{so}(3)$ at $t = 0$ and consider the derivative of the constraint at $t = 0$. The product rule yields

$$\frac{d}{dt}|_{t=0} R(t)^T R(t) = X^T + X = 0. \tag{A.108}$$

This implies that any element of $\mathfrak{so}(3)$ is a skew-symmetric matrix. The derivative of the determinant constraint is

$$\frac{d}{dt}|_{t=0}|R(t)| = \mathrm{tr}(X) = 0. \tag{A.109}$$

However, this is automatically satisfied by any skew-symmetric matrix.

A.7.2.1 Adjoint Representation

The Lie algebra $\mathfrak{so}(3)$ can be identified bijectively with $\mathbb{R}^3$ using the following mapping:

$$* : \mathfrak{so}(3) \to \mathbb{R}^3 \tag{A.110}$$

$$* \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}. \tag{A.111}$$

This star notation recalls the classic Hodge dual of differential forms and is in fact closely related to that mapping[6]. As such, we use the same $*$ to denote the inverse of this mapping.

Notice that the $*$ mapping is related to the vector cross product:

$$\begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \begin{pmatrix} d \\ e \\ f \end{pmatrix} = \begin{pmatrix} bf - ce \\ cd - af \\ ae - bd \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \times \begin{pmatrix} d \\ e \\ f \end{pmatrix}. \tag{A.112}$$

Using this, the form of the adjoint action in SO(3) is derived as follows. First, let $R \in \mathrm{SO}(3)$ and denote by $R_i \in \mathbb{R}^3, i = 1, 2, 3$ vectors representing the transposed *rows* of R. Then, letting $X \in \mathfrak{so}(3)$, the adjoint action of R on X is

$$\mathrm{Ad}_R X = RXR^T = \begin{pmatrix} R_1^T \\ R_2^T \\ R_3^T \end{pmatrix} X \begin{pmatrix} R_1 & R_2 & R_3 \end{pmatrix} \tag{A.113}$$

$$= \begin{pmatrix} R_1^T \\ R_2^T \\ R_3^T \end{pmatrix} \begin{pmatrix} *X \times R_1 & *X \times R_2 & *X \times R_3 \end{pmatrix} \tag{A.114}$$

$$= \begin{pmatrix} R_1 \cdot (*X \times R_1) & R_1 \cdot (*X \times R_2) & R_1 \cdot (*X \times R_3) \\ R_2 \cdot (*X \times R_1) & R_2 \cdot (*X \times R_2) & R_2 \cdot (*X \times R_3) \\ R_3 \cdot (*X \times R_1) & R_3 \cdot (*X \times R_2) & R_3 \cdot (*X \times R_3) \end{pmatrix}. \tag{A.115}$$

Now, using that triple products are invariant under even permutation of the arguments,

$$A \cdot (B \times C) = C \cdot (A \times B) = B \cdot (C \times A), \tag{A.116}$$

---

[6] If one identifies 1-forms with vector fields and 2-forms with skew symmetric matrix fields in $\mathbb{R}^3$, the Hodge dual represents exactly the mapping above at each point.

and that the cross product is alternating in its arguments,

$$A \times B = -B \times A, \tag{A.117}$$

this is rewritten:

$$\begin{aligned}
\mathrm{Ad}_R X &= \begin{pmatrix} 0 & -R_2 \cdot (*X \times R_1) & R_1 \cdot (*X \times R_3) \\ R_2 \cdot (*X \times R_1) & 0 & -R_3 \cdot (*X \times R_2) \\ -R_1 \cdot (*X \times R_3) & R_3 \cdot (*X \times R_2) & 0 \end{pmatrix} && \text{(A.118)} \\
&= * \begin{pmatrix} R_3 \cdot (*X \times R_2) \\ R_1 \cdot (*X \times R_3) \\ R_2 \cdot (*X \times R_1) \end{pmatrix} && \text{(A.119)} \\
&= * \begin{pmatrix} (R_2 \times R_3) \cdot *X \\ (R_3 \times R_1) \cdot *X \\ (R_1 \times R_2) \cdot *X \end{pmatrix} && \text{(A.120)} \\
&= * \left(\begin{pmatrix} (R_2 \times R_3)^T \\ (R_3 \times R_1)^T \\ (R_1 \times R_2)^T \end{pmatrix} (*X) \right). && \text{(A.121)}
\end{aligned}$$

Thus $* \mathrm{Ad}_R *$ can be represented by the 3×3 matrix

$$* \mathrm{Ad}_R * = \begin{pmatrix} (R_2 \times R_3)^T \\ (R_3 \times R_1)^T \\ (R_1 \times R_2)^T \end{pmatrix}. \tag{A.122}$$

Since R is an orthogonal matrix, its rows form an orthonormal ordered basis. This implies that

$$R_i = R_j \times R_k \tag{A.123}$$

where $\{ijk\}$ is an even permutation of $\{123\}$. Using that fact, the above matrix is seen to be simply the original matrix R:

$$* \mathrm{Ad}_R * = R. \tag{A.124}$$

The infinitesimal adjoint action can be computed using the earlier result that in matrix groups, $\mathrm{ad}_X Y$ is simply the matrix commutator $[X, Y]$. Computing the commutator for skew-symmetric matrices yields

$$\begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \begin{pmatrix} 0 & -f & e \\ f & 0 & -d \\ -e & d & 0 \end{pmatrix} = \begin{pmatrix} -cf-be & bd & cd \\ ae & -cf-ad & ce \\ af & bf & -be-ad \end{pmatrix} \tag{A.125}$$

$$\begin{pmatrix} 0 & -f & e \\ f & 0 & -d \\ -e & d & 0 \end{pmatrix} \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} = \begin{pmatrix} -cf-be & ae & af \\ bd & -cf-ad & bf \\ cd & ce & -be-ad \end{pmatrix} \tag{A.126}$$

$$\left[\begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}, \begin{pmatrix} 0 & -f & e \\ f & 0 & -d \\ -e & d & 0 \end{pmatrix} \right] = \begin{pmatrix} 0 & bd-ae & cd-af \\ ae-bd & 0 & ce-bf \\ af-cd & bf-ce & 0 \end{pmatrix}. \tag{A.127}$$

Under the star mapping, this is simply a cross product:

$$\mathrm{ad}_X Y = [X, Y] = *(*X \times *Y) \tag{A.128}$$

Just as $\mathfrak{so}(3)$ is mapped to $\mathbb{R}^3$ using the $*$ mapping, so can the dual space $\mathfrak{so}(3)^*$ be mapped. The dual pairing is then proportional to the scalar product of vectors[7]:

$$(\mu, X) = 2(*\mu)^T (*X). \tag{A.129}$$

As the dual pairing is a vector dot product, linear operators are conjugated by matrix transpose, resulting in the following expressions for the dual adjoint representation of SO(3) over $\mathfrak{so}(3)$:

$$*(\mathrm{Ad}^*_R \, \mu) = R^T (*\mu) \tag{A.130}$$

$$*(\mathrm{ad}^*_X \, \mu) = (X)^T (*\mu) = -X(*\mu) = -(*X \times *\mu). \tag{A.131}$$

Under the $*$ mapping, elements of the Lie algebra $\mathfrak{so}(3)$ and the Lie co-algebra $\mathfrak{so}(3)^*$ are represented by vectors in $\mathbb{R}^3$. A left or right invariant metric is equivalent to a choice of inner product on the Lie algebra; in this case, such a metric is simply described by a symmetric positive definite matrix A. The inner product between two Lie algebra elements is then

$$\langle X, Y \rangle = X^T A Y. \tag{A.132}$$

This, along with the Euler-Poincaré formulas, fully describe geodesics in SO(3).

---

[7]The factor of 2 comes from the repeated coefficients in a general skew-symmetric matrix, but can usually be ignored.

A.7.2.2 Biinvariant Metrics

Let $v \in T_R \operatorname{SO}(3)$ be a tangent vector at a group element R. The skew symmetric matrix $X = R^{-1}v \in \mathfrak{so}(3)$ is the left trivialization of v, while $Y = vR^{-1} \in \mathfrak{so}(3)$ is the right trivialization. Notice that

$$Y = \operatorname{Ad}_R X \tag{A.133}$$

$$*Y = R(*X). \tag{A.134}$$

Given a right invariant metric on SO(3), determined by an inner product matrix A, the metric written in terms of right trivialization is

$$\langle v, w \rangle_{T_R \operatorname{SO}(3)} = (*Y)^T A(*Y) \tag{A.135}$$

$$= (R(*X))^T AR(*X) \tag{A.136}$$

$$= (*X)^T R^T AR(*X). \tag{A.137}$$

This shows that, given a matrix A determining a right invariant metric on SO(3), the corresponding matrix for the inner product of left trivialized vectors is $R^T AR$. In particular, this means that a right invariant metric is only left invariant if, for all $R \in \operatorname{SO}(3)$,

$$R^T AR = A. \tag{A.138}$$

Thus for any biinvariant metric, A must commute with every element of SO(3).

Let x be an eigenvector of R. Every rotation matrix R has exactly one such eigenvector (up to scalar multiplication) representing the axis of rotation, whose eigenvalue is one. Applying the above formula to x,

$$R^T ARx = Ax \tag{A.139}$$

$$R^T Ax = Ax, \tag{A.140}$$

meaning Ax must be an eigenvector for R^T as well. However, as R and R^T are inverse to one another, they share eigenvectors. In particular, since in general they have only one eigenvector, this implies that

$$Ax = \lambda x. \tag{A.141}$$

Note that this is true for any x, since for each x there exists a rotation that fixes x.

To see that λ must be independent of x, let $x_1, x_2 \in \mathbb{R}^3$ be two noncollinear vectors, such that

$$Ax_1 = \lambda_1 x_1 \tag{A.142}$$

$$Ax_2 = \lambda_2 x_2. \tag{A.143}$$

From this follows

$$A(x_1 + x_2) = \lambda_1 x_1 + \lambda_2 x_2 \tag{A.144}$$

$$= \lambda_1 \left(x_1 + \frac{\lambda_2}{\lambda_1} x_2 \right). \tag{A.145}$$

Since $x_1 + x_2$ is also an eigenvector for A, it must be the case that $\lambda_1 = \lambda_2$. We conclude that any biinvariant metric on SO(3) must have a matrix A which is some scalar λ times the identity matrix.

The biinvariant metric on SO(3) is convenient as it allows for closed form computation of geodesics. Let $X \in \mathfrak{so}(3)$ represent an initial velocity vector, and consider the integrated form of the geodesic equation:

$$\frac{d}{dt} R(t) = R(t) \operatorname{Ad}^{\dagger}_{R(t)} X \tag{A.146}$$

$$= R(t) A^{-1} \operatorname{Ad}^{*}_{R(t)} AX \tag{A.147}$$

$$= R(t) \frac{1}{\lambda} \operatorname{Ad}^{*}_{R(t)} \lambda X \tag{A.148}$$

$$= R(t) \operatorname{Ad}^{*}_{R(t)} X \tag{A.149}$$

$$= R(t) R(t)^T X R(t) \tag{A.150}$$

$$= X R(t). \tag{A.151}$$

This is precisely the defining equation for the matrix exponential, implying that geodesics from the identity in SO(3) under the unique biinvariant metric have the following closed form solution:

$$R(t) = \exp(tX). \tag{A.152}$$

The form of these geodesics provides justification for the terminology "exponential map" and "log map", representing functions defined on Riemannian manifolds to compute geodesics given initial velocities and vice versa.

References

[1] V. I. Arnol'd. *Mathematical Methods of Classical Mechanics*. 2nd. Springer, 1989. ISBN: 0387968903.

[2] N. Bou-Rabee. "Hamilton-Pontryagin Integrators on Lie Groups". PhD thesis. California Institute of Technology, 2007.

[3] M. Bruveris et al. "The Momentum Map Representation of Images". In: *Journal of Nonlinear Science* 21.1 (2011), pp. 115–150.

[4] Francesco Bullo. "Invariant Affine Connections and Controllability On Lie Groups". In: (1995).

[5] Manfredo P do Carmo. *Riemannian Geometry*. 1st. Birkhäuser Boston, 1992.

[6] P Thomas Fletcher. "Geodesic Regression and the Theory of Least Squares on Riemannian Manifolds". In: *International Journal of Computer Vision* (2012), pp. 1–15.

[7] Darryl D Holm. *Geometric Mechanics: Dynamics and Symmetry. Part 1*. Vol. 1. Imperial College Press, 2008.

[8] SHOSHICHI Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry: Vol.: 1*. Wiley-Interscience, 1996.

[9] A.D. Lewis and R.M. Murray. "Configuration Controllability of Simple Mechanical Control Systems". In: *SIAM Journal on Control and Optimization* 35.3 (1997), pp. 766–790.

[10] J.E. Marsden and T.S. Ratiu. *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*. Vol. 17. Springer, 1999.

[11] L Younes, F Arrate, and M I Miller. "Evolutions Equations in Computational Anatomy". In: *NeuroImage* 45.1 (2009), S40–S50.

APPENDIX B

GROUP ACTIONS

Lie groups (see Appendix A) are useful because they represent *symmetry*, a term which hints that the symmetries they describe are those of some objects on which they operate. The adjoint representation gives a way to interpret Lie groups as transformations of their tangent spaces. This is an example of a group *acting* on a vector space. In general, a *Lie group action* of a Lie group G on a smooth manifold M is defined as a mapping that takes every group element g to an automorphism[1] of M, for which composition of automorphisms is compatible with the group action. The group action is denoted with a lower dot and these conditions are written

$$g.p \in M \qquad \forall g \in G, \quad p \in M \tag{B.1}$$

$$e.p = p \tag{B.2}$$

$$h.(g.p) = (hg).p \ . \tag{B.3}$$

In this appendix, I will touch the surface of the theory of geodesics and geodesic regression in Lie group actions. The reader is referred to the wide body of literature within geometric mechanics for a deeper treatment of this topic [5, 4, 3]. In particular, most of the content in this appendix is covered in the recent publication Bruveris et al. [2].

B.1 Infinitesimal Generators and Momentum Maps

Since G is a Lie group, this group action can be differentiated with respect to the group element. Once again, let $g(t)$ be a curve through the identity with velocity ξ at $t = 0$. The derivative of the curve $g(t).p \in M$ is a tangent vector in M at p called the *infinitesimal generator* of the group action, and is also written with a dot

$$\xi.p = \frac{d}{dt}\big|_{t=0} g(t).p \in T_pM. \tag{B.4}$$

[1] a diffeomorphism from M to itself

Note that this is not itself a group action but instead is a mapping $\xi. : M \to TM$ mapping each point in M to a tangent vector at that point.

Fixing the point $p \in M$, the mapping

$$\rho_p : \mathfrak{g} \to T_pM \tag{B.5}$$

$$\rho_p(\xi) = \xi.p \tag{B.6}$$

is a linear map from the Lie algebra of G to the tangent space at p. As such, its dual ρ_p^* exists and is a linear mapping from the cotangent space at p to the Lie coalgebra

$$\rho_p^* : T_pM^* \to \mathfrak{g}^*. \tag{B.7}$$

Dropping the subscript, this mapping is usually written as $\mathbf{J} : TM^* \to \mathfrak{g}^*$. Such a mapping takes a covector in the *object space* M and returns a *momentum* in the Lie coalgebra. Hence, $\mathbf{J}$ is called the *cotangent lift momentum map* [2]. This operation is analogous to the conversion of a tangential force applied at a point on the 2-sphere into an *angular momentum*, considered as an element of the Lie coalgebra of the rotation group SO(3).

A remarkable fact about $\mathbf{J}$ is that it is conserved by the coadjoint action Ad^*, in the following sense:

$$\mathrm{Ad}_g^* \mathbf{J}(p, m) = \mathbf{J}(g.p, gm). \tag{B.8}$$

Suppose $g(t)$ is a geodesic curve of group elements acting on a point p_0, so that at any time

$$p(t) = g(t).p_0. \tag{B.9}$$

If the momentum of the geodesic is originally *horizontal*, meaning it is of the form

$$\mu(0) = \mathbf{J}(p_0, m_0) \tag{B.10}$$

for some $m_0 \in T_{p_0}M^*$, then the momentum of $g(t)$ *remains* horizontal:

$$\mu(t) = \mathrm{Ad}_{g(t)}^* \mathbf{J}(p_0, m_0) = \mathbf{J}(g(t).p_0, g(t)m_0). \tag{B.11}$$

Typically, the momentum map is quite easy to compute, as is the group action, whereas the coadjoint action may be more complicated. As a result, this is often a reliable way to compute geodesics when considering group actions. Indeed, it is the basis for the *scalar momentum* approach to the diffeomorphic geodesic shooting [9].

B.2 Geodesic Regression for Linear Lie Group Actions

In this section, I will consider the case when the space M is a vector space V equipped with an inner product. I will also assume that G exploits the structure of V by acting linearly:

$$g.(av + bw) = ag.v + bg.w \qquad \forall g \in G, \quad a, b \in \mathbb{R}, \quad v, w \in v. \tag{B.12}$$

As the action of any element constitutes a linear mapping on V, it has a corresponding dual mapping, $g.^* : V^* \to V^*$. We define that as the action of the *inverse*, g^{-1}, on covectors in V^*:

$$(g^{-1}.\alpha, v) = (\alpha, g.v) \qquad \forall g \in G, \quad \alpha \in V^*, \quad v \in V. \tag{B.13}$$

The inverse is necessary so that the group action assumptions still hold. Note also that this dual action is infinitesimally generated by Lie algebra elements:

$$(-\xi.\alpha, v) = (\alpha, \xi.v) \qquad \forall \xi \in \mathfrak{g}, \quad \alpha \in V^*, \tag{B.14}$$

where this negative sign comes from the inverse in the dual group action definition.

The famous diamond notation [4, 2] is used to represent the cotangent lift momentum map: for any point $v \in V$ and covector $\alpha \in V^*$,

$$v \diamond \alpha = \mathbf{J}(v, \alpha). \tag{B.15}$$

In the remainder of this section, I will use this notation to derive explicit formulas for adjoint optimization of a least-squares regression problem, given data in V.

Suppose we are given a collection of vectors $y_i \in V$, we fix $g(0) = e$, and want to estimate $\xi(0) \in \mathfrak{g}$ and $v_0 \in V$ such that $g(t)$ is a geodesic and minimizes

$$E(\xi(0)) = \frac{1}{2} \sum_i \|g(t_i).v_0 - y_i\|^2. \tag{B.16}$$

The variation of this with respect to the curve point $g(t_i)$ is

$$g^{-1}(\delta_{g(t_i)} E) = (g.v_0) \diamond (g.v_0 - y_i)^\flat, \tag{B.17}$$

where the musical flat symbol, $\flat$, denotes the mapping $V \to V^*$ corresponding to the inner product on V, analogous to the inertia operator L. To see this, let $\delta g = Zg \in T_g G$ be a variation of $g \in G$. Then

$$(\delta_g E, \delta g)_{(T^*_g G, T_g G)} = \frac{1}{2}\delta \|g.v_0 - y\|_V^2 \tag{B.18}$$

$$= \frac{1}{2}\frac{d}{ds}\big|_{s=0} \| \exp(sZ) g.v_0 - y\|_V^2 \tag{B.19}$$

$$= ((g.v_0 - y)^\flat, \frac{d}{ds}\big|_{s=0} \exp(sZ) g.v_0)_{(V^*,V)} \tag{B.20}$$

$$= ((g.v_0 - y)^\flat, Zg.v_0)_{(V^*,V)} \tag{B.21}$$

$$= ((g.v_0 - y)^\flat, \rho_{g.v_0} Z)_{(V^*,V)} \tag{B.22}$$

$$= (\rho^*_{g.v_0}(g.v_0 - y)^\flat, Z)_{(\mathfrak{g}^*,\mathfrak{g})} \tag{B.23}$$

$$= ((g.v_0) \diamond (g.v_0 - y)^\flat, Z)_{(\mathfrak{g}^*,\mathfrak{g})}, \tag{B.24}$$

where I have subscripted the vector spaces for clarity.

These variations of E with respect to $\gamma(t_i)$ provide the jump discontinuities in the adjoint variable λ. Introducing a helper variable λ_i,

$$\lambda_i = \mathrm{Ad}^*_{g(t_i)} \lambda(t_i) \tag{B.25}$$

$$= \mathrm{Ad}^*_{g(t_i)} (g(t_i).v_0) \diamond (g(t_i).v_0 - y_i)^\flat \tag{B.26}$$

$$= v_0 \diamond (g^{-1}(t_i).(g(t_i).v_0 - y_i)^\flat). \tag{B.27}$$

Summing terms like this for all future times, the adjoint variable $\lambda(t)$ can be written

$$\lambda(t) = \sum_{t_i > t} \mathrm{Ad}^*_{g^{-1}(t)} \lambda_i \tag{B.28}$$

$$= \sum_{t_i > t} \mathrm{Ad}^*_{g^{-1}(t)} v_0 \diamond (g^{-1}(t_i).(g(t_i).v_0 - y_i)^\flat) \tag{B.29}$$

$$= \sum_{t_i > t} (g(t).v_0) \diamond (g(t) g^{-1}(t_i).(g(t_i).v_0 - y_i)^\flat). \tag{B.30}$$

This shows that not only does the first adjoint variable need not be integrated, but it is horizontal at all times, and its computation requires only the diamond map and the group action. In particular, this underlies the famous result that optimal geodesics for image matching with diffeomorphisms must necessarily have initial momenta which are equal to a scalar measure times the initial image gradient [2].

A special case arises when G acts by isometries, meaning

$$\langle g.v, g.w \rangle = \langle v, w \rangle, \qquad \forall g \in G, \quad v, w \in V. \tag{B.31}$$

In that case, the flatting operation commutes with the group action, and we have the following simplification [2]:

$$\lambda(t) = \sum_{t_i > t} (g_{0t}.v_0) \diamond (g_{t_i t} g_{0t_i}.v_0 - g_{t_i t}.y_i)^\flat) \tag{B.32}$$

$$\lambda(t) = \sum_{t_i > t} (g_{0t}.v_0) \diamond (g_{0t}.v_0 - g_{t_i t}.y_i)^\flat, \tag{B.33}$$

where I have used the shorthand

$$g_{st} = g(t)g(s)^{-1}, \tag{B.34}$$

to represent the group action from one time to another.

B.2.1 Optimal Template Estimation

Generally, the deforming template object v_0 will need to be estimated as well. In order to do so, observe that the variation with respect to v_0 is

$$\delta_{v_0} E = \sum_i g(t_i)^*.(g(t_i).v_0 - y_i)^\flat \tag{B.35}$$

$$= \sum_i g(t_i)^{-1}.(g(t_i).v_0 - y_i)^\flat. \tag{B.36}$$

This is the standard formula for variations with respect to a quadratic form in linear least squares optimization.

The variation with respect to v_0 could be used in a gradient descent scheme, but it is commonly solved in closed form by setting the above equation equal to zero. The steps involved are

$$\sum_i g(t_i)^{-1}.(g(t_i).v_0)^\flat = \sum_i g(t_i)^{-1}.y_i^\flat \tag{B.37}$$

$$\sum_i g(t_i)^{-1}.Ag(t_i).v_0 = \sum_i g(t_i)^{-1}.y_i^\flat \tag{B.38}$$

$$\left(\sum_i g(t_i)^{-1}.Ag(t_i). \right) v_0 = \sum_i g(t_i)^{-1}.y_i^\flat \tag{B.39}$$

$$v_0 = \left(\left(\sum_i g(t_i)^{-1}.Ag(t_i). \right)^{-1} \sum_i g(t_i)^{-1}.y_i^\flat \right)^\sharp. \tag{B.40}$$

Note that a few liberties were taken with notation in the lines above. For instance, I used the operator $A : V \to V^*$ to denote the flatting operation and $\sharp$ to denote its inverse, as is common. Also, there is no guarantee that the sum of linear operators will be invertible. Still, when this equation is able to be computed, it gives the optimal template in closed form, as in Singh et al. [7].

Note again that when G acts by isometries, this formula is further simplifed. The variation with respect to v_0 in that case is

$$\delta_{v_0} E = \sum_i (v_0 - g(t_0)^{-1}.y_i)^\flat, \tag{B.41}$$

which is easily solved for v_0:

$$\hat{v}_0 = \frac{1}{N} \sum_i (g(t_i)^{-1}.(y_i)^\flat)^\sharp = \frac{1}{N} \sum_i g(t_i)^{-1}.y_i. \tag{B.42}$$

B.3 Diffeomorphic Image Deformation

As the fundamental example in this dissertation, I now give the necessary details to use the preceding results in the context of diffeomorphic image warping. The diffeomorphism group $\mathrm{Diff}(\Omega)$ of some image domain Ω (typically a bounded convex subset of $\mathbb{R}^d$, consists of smooth invertible mappings $\Omega \to \Omega$, with the group operation simply composition of these mappings. The Lie algebra $\mathfrak{g}$ consists of smooth vector fields on Ω, and its dual, $\mathfrak{g}^*$ consists of *vector-valued measures*. The following are the adjoint operations [8]:

$$\mathrm{Ad}_\varphi v = (D\varphi \circ \varphi^{-1}) v \circ \varphi^{-1} \tag{B.43}$$

$$\mathrm{ad}_\xi \eta = -[\xi, \eta] = D\xi\eta - D\eta\xi \tag{B.44}$$

$$\mathrm{Ad}^*_\varphi m = D\varphi m \circ \varphi \tag{B.45}$$

$$\mathrm{ad}^*_\xi m = (D\xi)^T m + Dm\xi + m\nabla \cdot \xi \tag{B.46}$$

where the last formula only holds if m is differentiable (that is, if Dm exists) and is otherwise understood to hold in the weak sense. In $\mathrm{Diff}(\Omega)$, a right invariant metric is generally used, due to the Eulerian fluid interpretation of such metrics, and the invariance to particle relabelling [1].

The vector space that $\mathrm{Diff}(\Omega)$ acts on depends on the context, but is commonly the space of square-integrable real-valued images $V = L^2(\Omega)$. The dual, V^*, to this space is generally a space of scalar measures. The group action in this case is given by composition,

$$\varphi.I = I \circ \varphi^{-1}. \tag{B.47}$$

The inverse on the right-hand side may be surprising, but its presence results in an image that depicts structures moving along the trajectories described by φ itself. Using the group

action and the Taylor expansion, assuming I is differentiable, the action is infinitesimally generated by advection:

$$v.I = -\nabla I \cdot v. \tag{B.48}$$

From this one derives that the diamond map is simply the mapping from scalar to vector momenta given by multiplication of the scalar measure by the gradient of the image:

$$I \diamond \alpha = -\alpha \nabla I, \qquad \forall \alpha \in V^*, \quad I \in V. \tag{B.49}$$

From these relations, the entire theory of scalar momentum geodesic shooting [6] is derived by substitution into the results of the previous sections (c.f. [2] for more details).

References

[1] Vladimir I Arnol'd. "Sur la Géométrie Différentielle des Groupes de Lie de Dimension Infinie et ses Applications à l'Hydrodynamique des Fluides Parfaits". In: *Ann. Inst. Fourier* 16 (1966), pp. 319–361.

[2] M. Bruveris et al. "The Momentum Map Representation of Images". In: *Journal of Nonlinear Science* 21.1 (2011), pp. 115–150.

[3] Darryl D Holm. *Geometric Mechanics: Dynamics and Symmetry. Part 1.* Vol. 1. Imperial College Press, 2008.

[4] Darryl D Holm, Jerrold E Marsden, and Tudor S Ratiu. "The Euler-Poincaré Equations and Semidirect Products with Applications to Continuum Theories". In: *Adv. in Math.* 137 (1998), pp. 1–81.

[5] J.E. Marsden and T.S. Ratiu. *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems.* Vol. 17. Springer, 1999.

[6] Michael I. Miller, Alain Trouvé, and Laurent Younes. "Geodesic Shooting for Computational Anatomy". In: *Journal of Mathematical Imaging and Vision* 24.2 (2006), pp. 209–228. DOI: 10.1007/s10851-005-3624-0.

[7] Nikhil Singh et al. "A Vector Momentum Formulation of Diffeomorphisms for Improved Geodesic Regression and Atlas Construction". In: *International Symposium on Biomedical Imaging (ISBI)*. Apr. 2013.

[8] Alain Trouvé and Laurent Younes. "Metamorphoses Through Lie Group Action". In: *Foundations of Computational Mathematics* 5.2 (2005), pp. 173–198.

[9] L Younes, F Arrate, and M I Miller. "Evolutions Equations in Computational Anatomy". In: *NeuroImage* 45.1 (2009), S40–S50.

APPENDIX C

STOCHASTIC DOSE QUANTIFICATION

In addition to the validation studies presented in Chapter 2, an application of 4D MAP CT image reconstruction was published by Geneser et al. In that work, the motion fields computed using 4D MAP image reconstruction were used to convert stochastic breathing patterns into stochastic anatomical motion. In turn, that stochastic motion was used to compute mean dose delivery during radiation therapy, along with standard deviations quantifying the *uncertainty* in dose delivered to each point in the patient's anatomy. This appendix includes a reprint of the publication Geneser et al. (2011).

Reprinted from Medical Image Analysis 15 (2011) 640–649 with permission from Elsevier

Medical Image Analysis 15 (2011) 640–649

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Quantifying variability in radiation dose due to respiratory-induced tumor motion

S.E. Geneser [a,*], J.D. Hinkle [a], R.M. Kirby [a], B. Wang [b], B. Salter [b], S. Joshi [a]

[a] Scientific Computing & Imaging Institute, University of Utah, Salt Lake City, UT, USA
[b] Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA

ARTICLE INFO

Article history:
Available online 14 July 2010

Keywords:
Respiratory-induced dose variation
Stochastic modeling
Polynomial chaos
Stochastic collocation
Stereotactic body radiation therapy

ABSTRACT

State of the art radiation treatment methods such as hypo-fractionated stereotactic body radiation therapy (SBRT) can successfully destroy tumor cells and avoid damaging healthy tissue by delivering high-level radiation dose that precisely conforms to the tumor shape. Though these methods work well for stationary tumors, SBRT dose delivery is particularly susceptible to organ motion, and few techniques capable of resolving and compensating for respiratory-induced organ motion have reached clinical practice. The current treatment pipeline cannot accurately predict nor account for respiratory-induced motion in the abdomen that may result in significant displacement of target lesions during the breathing cycle. Sensitivity of dose deposition to respiratory-induced organ motion represents a significant challenge and may account for observed discrepancies between predictive treatment plan indicators and clinical patient outcomes.

Improved treatment-planning and delivery of SBRT requires an accurate prediction of dose deposition uncertainties resulting from respiratory motion. To accomplish this goal, we developed a framework that models both organ displacement in response to respiration and the underlying random variations in patient-specific breathing patterns. Our organ deformation model is a four-dimensional maximum a *posteriori* (MAP) estimation of tissue deformation as a function of chest wall amplitudes computed from clinically obtained respiratory-correlated computed tomography (RCCT) images. We characterize patient-specific respiration as the probability density function (PDF) of chest wall amplitudes and model patient breathing patterns as a random process. We then combine the patient-specific organ motion and stochastic breathing models to calculate the resulting variability in radiation dose accumulation. This process allows us to predict uncertainties in dose delivery in the presence of organ motion and identify tissues at risk of receiving insufficient or harmful levels of radiation.

1. Introduction

External beam radiotherapy destroys cancer cells by delivering ionizing radiation to a tumor. Because conventional radiation treatment delivers only a few unmodulated straight-line beams per treatment, the volume receiving radiation cannot be easily matched to the irregular shape of most tumors. Modern techniques like dynamic conformal arc and intensity modulated radiation therapy (IMRT) modulate the intensity or shape of external beams applied over many different angles to enable radiation dose delivery that precisely conforms to a physician-defined tumor geometry (Purdy, 2001). Combined with improved image guidance techniques that allow clinicians to identify tumor shapes and locations with greater accuracy (Xing et al., 2006), IMRT enables precise dose conformity to the targeted tumor volume (as demonstrated in Fig. 1). This process allows safe delivery of extremely large ablative radiation doses that dramatically increases the likelihood of tumor control (Timmerman et al., 2005) and reduces the collateral damage to surrounding healthy tissue, particularly in cases where the tumor is stationary during treatment. Hypo-fractionated stereotactic body radiation therapy (SBRT) combines conformal therapy and image guidance techniques to apply high levels of radiation over a few treatments (each treatment delivers a fraction of the total prescribed dose) and has proven safe and highly effective for controlling tumors of the lung, liver, and spine (McGarry et al., 2005).

State-of-the-art commercial treatment-planning systems generally calculate dose delivery distributions for static tissues within 3% accuracy (Siantar et al., 2001; Heath et al., 2004; Herk, 2004; Rassiah-Szegedi et al., 2006), but cannot yet calculate accurate dose in the presence of organ motion. Though respiratory-induced organ motion can result in significant movement during the breathing cycle (Lujan et al., 1999; Brandner et al., 2006) (as evidenced in Fig. 2), clinical radiation dose SBRT plans deliver dose to a static volume over all treatments and do not dynamically adjust to changing tumor position. Due to high dose gradients inherent in conformal radiation delivery, IMRT is particularly

* Corresponding author. Present address: Department of Radiation Oncology, Stanford University, Stanford, CA, USA.
E-mail address: sgeneser@stanford.edu (S.E. Geneser).

1361-8415/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2010.07.003

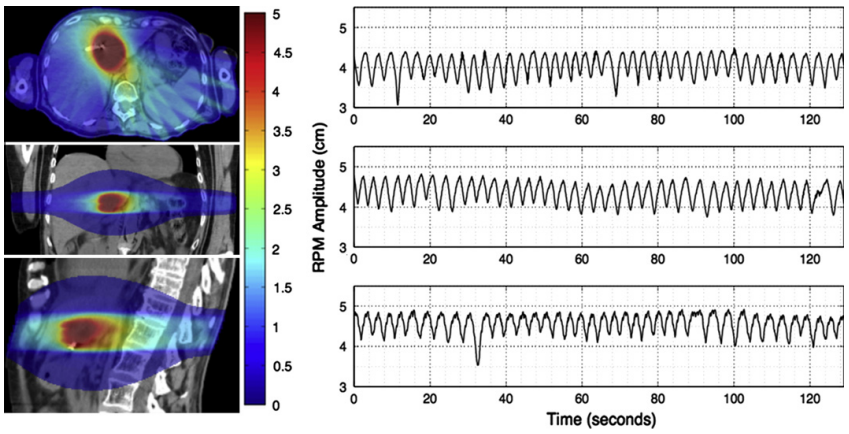


Fig. 1. *The planned (static) dose distribution and Real-time Position Management™ (RPM) traces* for the analyzed SBRT liver cancer patient illustrates the high spatial gradients of target-conforming dose and daily variations in breathing. The static deposited dose (in units of Gray) is color-mapped and superimposed on anatomical images for (from top to bottom) axial, sagittal, and coronal views. The RPM breathing traces are recorded for the same patient and time interval on different treatment days.

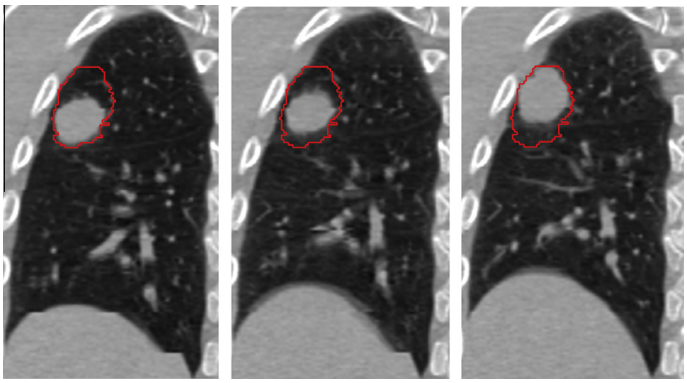


Fig. 2. *Respiratory-induced organ motion* can cause significant tumor displacement. These coronal slices depict recorded anatomy corresponding to three phases (from left to right: full-inhale, mid-cycle, and full-exhale) within the breathing cycle for a typical lung SBRT patient. The red outline denotes the physician defined internal target volume (ITV). Note the displacement of the tumor within the (stationary) clinical ITV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

susceptible to targeted tumor motion and respiration can lead to significant dose delivery errors (Bortfeld et al., 2002; Lujan et al., 2003; Chui et al., 2003; Jiang et al., 2003; Bortfeld et al., 2004; Vedam et al., 2005). The low number of treatment fractions renders SBRT even more sensitive to intra-fraction motion and studies investigating the dosimetric consequences of respiratory-induced tissue motion on SBRT have found variations between planned and delivered dose distributions as significant as 20% (Wu et al., 2008). The uncertainties resulting from respiratory-induced tissue motion complicate SBRT treatment of extracranial lesions and may well account for the observed discrepancies between predictive indicators and clinical patient outcome statistics. No process has been developed to accurately predict uncertainties in dose delivery resulting from random patient breathing patterns.

Controlling patient breathing during treatment and restricting beam-on times to windows of low variation in patient anatomy can reduce the variability in dose deposition due to respiratory-induced motion. Specific methods to reduce dose variation include respiratory-gating (Keall et al., 2005b), breath-hold (Hanley et al., 1996), and coached breathing (Neicu et al., 2006). However, each has limitations and none are appropriate for all patients (Keall et al., 2006). For example, breath-hold techniques can induce an unacceptable level of patient discomfort (particularly for lung cancer patients with severely compromised respiratory function), respiratory-gating significantly increases the treatment time because it restricts beam delivery to a small fraction of the treatment period, and coached breathing may be impractical because some patients are not trainable. While some radiation oncologists employ these methods, most instead design a treatment based on more fundamental mechanisms, e.g., the inclusion of a border or margin around the defined internal target volume (ITV). This widely employed technique is intended to accommodate tumor motion during treatment and ensure sufficient dose delivery by treating a "motion envelope" that encompasses the tumor positions observed during breathing. However, tumor volumes and margins for treatment-planning are generated from images ob-

tained on a single day that cannot be used to predict the subsequent variations in breathing. Moreover, this approach ensures a complete treatment of the target at the expense of irradiating adjacent healthy tissues. Alternative treatment methods have been developed that adjust to or move with the tumor in order to accommodate respiratory-motion. Such methods are successful in some cases but cannot yet be widely applied. For example, automated tumor tracking and delivery (Keall et al., 2004; Naqvi et al., 2005; Sawant et al., 2008) works well for lung tumors that are clearly discernible in CT and X-ray images, but typically not for liver, which can be difficult to distinguish from surrounding tissue. As a result, it is challenging to implement robust and accurate automatic tracking methods for liver tumors. For thorax and abdominal tumors, breathing motion remains one of the major obstacles to reducing the irradiation volume while maintaining a high probability of treatment success (Jiang et al., 2008).

Patient breathing is not time-periodic (or perfectly repeatable), and respiration patterns can vary significantly between treatment fractions. The fundamentally random fluctuations in respiratory-induced organ motion can result in delivered doses that significantly vary from treatment to treatment. Failure to accommodate patient breathing motion randomness can result in under-dosing of the target and/or deposition of potentially dangerous dose levels to surrounding healthy tissue. When limiting patient organ motion during treatment is impossible or unreasonable, it is essential to incorporate an accurate prediction of the effects of the stochastic respiratory process on dose deposition for improved safety and efficacy of SBRT treatment-planning and delivery. Though several groups have worked to develop accurate models that incorporate the effect of respiratory-induced organ motion on dose deposition (Boldea et al., 2008), no commercially available computational tools successfully address this problem. We propose an approach capable of predicting the variance in dose accumulation for SBRT treated abdominal lesions resulting from stochastic organ motion induced by variations in patient breathing patterns.

We apply our framework (described previously in Geneser et al. (2009)) to quantify the impact of variations in patient-specific breathing patterns on dose deposition for a typical SBRT liver patient. The anatomical CT images, clinical dose plans, and forward dose calculations used in this work were obtained during the Huntsman Cancer Institute's clinical planning and treatment process. The patient's static dose treatment plan and breathing traces from three treatment days are depicted in Fig. 1. We provide a flow chart (depicted in Fig. 3) to outline the major components of our procedure and indicate how the clinical data is incorporated into our framework. While the results presented here are a retrospective analysis, the same approach can be applied to predict dose uncertainties on a patient-specific basis prior to treatment with minimal alteration of the clinical planning routine. Because the anatomical CT images are collected and dose distributions are calculated as part of a typical SBRT dose planning process, our framework requires only that breathing traces be obtained on a few days prior to treatment. This can be accomplished without extending the planning time because the dose optimization process currently requires several days of computation time during which the breathing traces can be recorded and analyzed.

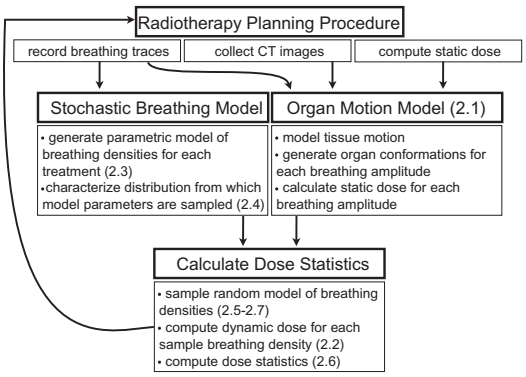


Fig. 3. *Overview* of the procedure for calculating variations in radiation dose resulting from fluctuations in respiratory-induced organ motion. Where appropriate, the sections describing the components are denoted.

To predict the variability in radiation dose delivery resulting from random patient breathing patterns, we build a model of patient-specific respiratory-induced organ motion to compute the dynamic dose deposition in response to recorded breathing behavior during a given treatment using the method described in Hinkle et al. (2009). We then model patient-specific breathing patterns as a stochastic process by parametrizing the recorded breathing traces and modelling the resulting breathing parameters as random variables. Once we estimate the underlying distributions of the random variables, we incorporate our stochastic breathing model into the dynamic dose computation that accounts for variations in organ motion during treatment.

Monte Carlo techniques cannot be employed to solve such systems because the large number of solutions necessary to converge to accurate statistics and long computational times required to generate a single dynamic dose solution renders such methods infeasible. Using polynomial chaos (Wiener, 1938; Xiu and Karniadakis, 2002) and Smolyak collocation (Mathelin and Hussaini, 2003; Babuška et al., 2005; Xiu and Hesthaven, 2005; Xiu, 2007; Ganapathysubramanian and Zabaras, 2007; Nobile et al., 2008) techniques significantly reduces the number of dynamic dose calculations and thus the cost of computing accurate dose statistics. Indeed, the speedup is significant enough to render incorporating the framework into the clinical optimization and planning process feasible. Using this method, we compute pertinent dose statistics to predict and assess variations in radiation dose due to random variations in patient breathing patterns subsequent to the clinical planning process. Our goal is to enable clinicians to identify SBRT dose plans that are robust to fluctuations in patient respiratory patterns and improve tumor control and normal tissue sparing.

2. Methods

To account for stochastic respiratory-induced tumor motion, we first quantify the impact of organ motion on dose delivery over the course of a treatment. Calculating the dynamic dose requires both an accurate patient-specific anatomical motion model and the ability to calculate static dose deposition at anatomical configurations observed during unrestricted patient breathing. Commercially available four-dimensional respiratory-correlated computed tomography (4D RCCT) (Ford et al., 2003; Vedam et al., 2003) tools provide a means of visualizing four-dimensional organ motion, and clinicians currently rely on the detailed images produced from such scans to generate the volume contours to be irradiated. Using deformable image registration techniques, the anatomical CT configurations observed during the breathing cycle can be mapped onto a common geometry. This mapping can be used to compute the dynamic dose accumulation resulting from observed or simulated respiratory-induced tissue deformations (Foskey et al., 2005; Keall et al., 2005a).

We build an explicit model of tissue deformation from anatomical CT patient images obtained during the breathing cycle. Our motion model and subsequent dose calculations rely on the well-

justified and widely accepted assumption that the anatomical configuration is a function of breathing pattern as measured by a surrogate signal, e.g., the real-time position management (RPM™) system (Kubo et al., 2000) (Varian RPM, Varian Medical Systems Inc., Palo Alto, CA). Several groups have investigated the relationship between external and internal motion markers (Beddar et al., 2007; Ionascu et al., 2007) and reported high correlation between the two. This correspondence is the basis for the respiratory-correlated spiral CT (RCCT) method that generates CT images at multiple respiratory phases from a single spiral CT scan (Ford et al., 2003; Vedam et al., 2003).

From the patient images, we construct a deformation field, $h(\vec{x}, a(t))$, that maps each spatial point, $\vec{x}$, in a base image to its corresponding anatomical position as a function of breathing amplitude. Generating deformation fields that model organ motion is a well-studied problem and several groups have developed techniques that produce accurate deformations from artifact-free CT images (Pevsner et al., 2006; Wijesooriya et al., 2008). We model the three-dimensional tissue motion as a function of breathing amplitude rather than phase. Several groups have shown that organ position correlates highly with breathing amplitude (Nehmeh et al., 2004; Chi et al., 2006; Abdelnour et al., 2007). In the following sections, we describe our methods to construct a set of amplitude-indexed deformations which represent the fully deformable diffeomorphic transformation of a base image in response to breathing motion.

2.1. Four-dimensional geometric model of organ deformation

To construct a physiologically realistic model of organ motion, we follow the 4D maximum a *posteriori* (MAP) image reconstruction approach proposed by Hinkle et al. (2009) that estimates amplitude-varying velocity fields acting on the tissue during breathing from 4D RCCT images. Rather than calculating correspondences between pairs of binned images, this method simultaneously estimates deformations and the base organ configuration over the entire set of collected CT data. In contrast, binning methods discard a significant amount of the CT data and can result in image artifacts due to mismatched slices. In this framework, one estimates a 4D time-indexed image, $I_t(\vec{x})$, by maximizing a *posterior* likelihood that combines prior motion information with a data log-likelihood term derived from a noise model.

The raw data are modeled as a set of projections, $d_i \in L^2(\Omega_d)$, obtained via linear operators $P_i{:}L^2(\Omega) \to L^2(\Omega_d)$, where $\Omega \subset \mathbb{R}^3$ is the image domain, and Ω_d is the data domain. In this notation, $L^2(X)$ denotes the Hilbert space of square-integrable functions over the domain, X, equipped with the usual inner product. For conebeam CT, the projection operator is the conebeam projection and $\Omega_d \subset \mathbb{R}^2$ is two-dimensional because the detector consists of a two-dimensional array of elements. In the case of fan beam images, the data domain, $\Omega_d \subset \mathbb{R}^1$, is one-dimensional and the operator is given by the Radon transform. Because organ motion is considerably slower than RCCT gantry rotation, we reconstruct each slice assuming no anatomical motion and P_i becomes a simple slice selection operator.

Deriving an expression for the data log-likelihood requires a model of the noise characteristics of the scanner. For CT data with sufficiently high signal-to-noise ratio, the noise is approximately Gaussian, and the data log-likelihood is a sum of squared differences between the projected image estimate and the data as follows:

$$L(\{d_i\}|I_t) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}\int_{s\in\Omega_d}|P_i\{I_{t_i}\}(s) - d_i(s)|^2 ds. \tag{1}$$

It is worth noting that interpreting Eq. (1) as a log-likelihood function is non-trivial. Extending the notion of a normal distribution to the infinite-dimensional Hilbert space, $L^2(\Omega)$, in a rigorous way is possible but requires careful treatment of stochastic processes and Gaussian random fields. For a more in-depth discussion of these issues see Christensen et al. (1996) and Dupuis et al. (1998).

We model the 4D image, $I_t(\vec{x}) = I_0 \circ g(\vec{x}, t)$, as a 3D base image, I_0, undergoing a time-indexed deformation, $g(\vec{x}, t)$. Assuming organ motion is correlated with breathing amplitude, the deformations are amplitude-indexed as $h(\vec{x}, a(t))$. The velocity of a point, $\vec{x}$, in the patient's anatomy is described by the ordinary differential equation,

$$\frac{d}{dt}h(\vec{x}, a(t)) = v(h(\vec{x}, a(t)), a(t))\frac{da}{dt}, \tag{2}$$

where $v(h(a, \vec{x}), a) = \frac{\partial}{\partial a}h(\vec{x}, a)$ is a velocity field indexed by breathing amplitude rather than time. The deformation from the base amplitude is given by the associated integral equation,

$$h(\vec{x}, a) = \vec{x} + \int_0^a v(h(\vec{x}, a'), a')da'. \tag{3}$$

This formulation guarantees that the resulting estimates of patient anatomy are diffeomorphic to one another and ensures that organs do not tear or disappear during breathing (Joshi et al., 2000). Diffeomorphic deformations also provide a one-to-one correspondence between image points, which enables tissue trajectory tracking. We enforce smoothness by introducing a prior on the velocities via a Sobolev norm, $\|v\|_V^2$, defined by:

$$\|v\|_V^2 = \langle v, v\rangle_V = \int_0^1\int_{\vec{x}\in\Omega}\|Lv(\vec{x}, a)\|_{\mathbb{R}^3}^2 d\vec{x}da, \tag{4}$$

where L is a differential operator chosen to reflect physical tissue properties. Following Kuo, 1975 we place a Gaussian prior on the Sobolev space, which is embedded in a Banach space of continuous vector fields. The continuity properties of elements in the Banach space are determined by the choice of Sobolev space, which in turn is determined by the choice of differential operator L. In our implementation, $Lv = -\alpha\nabla^2 v + \gamma v$ for scalar parameters α and γ, following Christensen et al. (1996, 2005).

We enforce further physical tissue properties by constraining the velocity fields. In particular, if the divergence of the velocity field is zero, the resulting deformation has unit Jacobian determinant and is locally volume preserving. This is a necessary constraint when modeling the breathing induced motion of incompressible fluid-filled organs such as liver.

Combining the data log-likelihood with the motion prior, the log-*posterior* likelihood of observing the data takes the form:

$$\mathcal{L}(I_0, v|d_i) = -\|v\|_V^2 - \frac{1}{2\sigma^2}\sum_i\int_{s\in\Omega_d}|P_i\{I_0 \circ h(\vec{x}, a_i)\}(s) - d_i(s)|^2 ds. \tag{5}$$

A MAP estimate that maximizes Eq. (5) with respect to both the base image and deformation is obtained via an alternating iterative algorithm that updates the deformation and image estimates at each iteration using a gradient ascent step and the associated Euler–Lagrange equation, respectively. Following the approach of Beg et al. (2005), efficient computations of the Euler–Lagrange equations are implemented in the Fourier domain, requiring only a matrix multiplication and Fourier transforms of v_k at each iteration of the algorithm. We enforce the zero-divergence velocity field constraint at each step in the Fourier domain, and additional implementation details can be found in Hinkle et al. (2009).

As with the data log-likelihood term, one must be careful when interpreting Eq. (5) as a posterior likelihood. A formal prior distribution may be placed on the space of velocity fields by defining a Gaussian random field on the Sobolev space characterized by the differential operator L. Because probability density functions do

not exist on the infinite-dimensional spaces in which we define our probability distributions, strict interpretation of the data and prior expressions as density functions is imprecise. However, an extension of MAP estimation to infinite-dimensional posterior distributions can be made precise. In this approach densities are replaced by the limits of probabilities of balls around a given point. In the case of a Gaussian random field, the result is an expression known as the Onsager–Machlup functional, which takes the form of an exponentiated squared norm, as is the case for both the data log-likelihood and prior terms in Eq. (5). As discussed by Dupuis et al. (1998), such a treatment is quite involved, and it is often more convenient to simply view the proposed approach as a minimum-energy estimation problem.

2.2. Incorporating organ motion into dynamic dose calculation

The dynamic dose deposition, D, accounts for the effects of known organ motion during a single treatment interval and is integrated over the time interval $[0,T]$ as follows:

$$D = \int_0^T d_t(g(\vec{x},t),t)dt. \tag{6}$$

The term, $d_t(g(\vec{x},t),t)$, is the time-dependent static dose over the patient's anatomy at time t, $g(\vec{x},t)$. A change of variables yields the total deposited dose over a treatment period as an integral over the amplitudes,

$$D = \int_{min(a)}^{max(a)} d_a(h(\vec{x},a),a)f(a)da, \tag{7}$$

where $d_a(h(\vec{x},a),a)$ is the amplitude-dependent dose corresponding to $a(t)$, the amplitude of the breathing signal during treatment, mapped to the base image according to the deformation field, $h(\vec{x},a(t))$, and $f(a)$ is the relative time density of the breathing amplitudes over the treatment interval. Given a set of amplitude-binned CT images and a model of the organ deformation as described above, we estimate the delivered dose by discretizing Eq. (7) to obtain a weighted sum of amplitude-indexed dose images as follows:

$$D = \sum_{i=0}^{N} w_i d(h(a_i),a_i), w_i = \int_{a_i-\delta a}^{a_i+\delta a} f(a)da, \tag{8}$$

where $\delta a = \frac{1}{2}(a_{i+1} - a_i)$ is the size of the amplitude discretization. The term $d(h(a_i),a_i)$ is the dose deposited to the tissues at the anatomical conformation corresponding to the breathing amplitude, a_i. The weights, w_i, account for the relative amount of time the breathing amplitude falls within the interval, $[a_i - \delta a, a_i + \delta a]$, during a treatment period.

It is important to stress that the dynamic model of dose deposition presented above accounts only for the respiratory-induced organ motion observed during a single treatment. As such, it includes none of the expected variability due to patient breathing motion. For D to provide insight to the effects of motion variability, one must incorporate a model of patient breathing variability into the dynamic dose deposition calculation. In the following sections, we provide the framework to characterize the stochastic nature of daily breathing patterns and apply our model to determine the resulting uncertainties in SBRT dose accumulation.

2.3. Parametrization of breathing amplitude density

The extent of breathing variability differs over individuals, necessitating patient-specific respiratory models to generate accurate predictions of radiation dosing resulting from random fluctuations in breathing patterns. Because the time density of breathing amplitudes is sufficient to accurately calculate a dose distribution over a treatment interval, we need only determine the variations in amplitude density as a function of time. To characterize the distributions from which a patient's breathing amplitude density is sampled on any given day, we first parametrize breathing density by fitting each patient breathing trace to a reasonable probability distribution. Parametrization is necessary to estimate the underlying distributions from which patient-specific amplitude densities are sampled and, ultimately, to develop a model that accurately captures the random fluctuations in patient breathing patterns. Once we fit the breathing densities, we then characterize the distributions of the parameters by performing principle component analysis. In this manner, we construct a patient-specific stochastic model of breathing that can be incorporated into the dose calculation to compute variances in dose deposition.

Gaussian Mixture Models (GMMs) provide a means of parametrizing the probability density of a random process (McLachlan and Peel, 2000) and are used here to model the amplitude density of individual RPM breathing traces. Such models are convex combinations of M Gaussian distributions as follows:

$$m(x,p_i,\mu_i,\sigma_i) = \sum_{i=1}^{M} p_i \frac{1}{\sigma_i\sqrt{2\pi}} e^{\frac{-(x-\mu_i)^2}{2\sigma_i^2}} \tag{9}$$

where μ_i and σ_i are the mean and standard deviation of the ith Gaussian distribution and p_i are positive weighting factors that sum to one. We fit these parameters to patient RPM breathing traces using the Expectation Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Peel, 2000) that optimizes the log-likelihood estimates of the GMM fits to the RPM breathing amplitude data. Because patients pause at inhale and exhale and the amplitudes for both are typically consistent over time, one often observes peaks in the amplitude density function at both locations. As a consequence, a two-Gaussian mixture model appears sufficient for estimating and parametrizing the amplitude density of RPM breathing traces of many patients. However, breathing patterns can differ widely among patients. For certain cases (particularly for moderate to advanced lung cancer patients with compromised and erratic respiratory status), breathing patterns may be more variable in frequency, rhythm, and depth of respiration, necessitating the use of GMMs with three or more Gaussians. The appropriate model should be chosen on a case-by-case basis. Given the parameters of the RPM amplitudes, we can analyze the characteristics of variability in the parameters for each patient and build a model to capture the patient-specific variations in daily breathing amplitude densities.

2.4. Model of breathing variability

After fitting breathing amplitude densities to Eq. (9), we estimate the variation of the parameters over observation days to characterize the patient's breathing fluctuations. Because it is only clinically feasible to obtain a small number of breathing traces (typically less than seven are collected per patient), it is difficult to generate accurate estimates of the underlying patient-specific distribution from which the GMM breathing parameters are sampled on any given day. It is important to note that the GMM parameters; p_i, μ_i, and σ_i, exhibit strong correlation. For example, one often observes a consistent distance between inhale and exhale amplitudes. This results in a high correlation between μ_i over the observation days. Using principal component analysis (PCA) (Pearson, 1901) we perform a linear transformation of the GMM parameters to identify the modes of greatest variation. We then formulate the patient-specific random GMM model parameters as a function of independent and uncorrelated Gaussian random variables, $\vec{\xi} = (\xi_1,\ldots,\xi_d)$, where d is the number of principal components (and thus random dimensions) necessary to accurately capture the breathing variability. The Gaussian random variables

are centered at zero and have variance corresponding to the eigenvalues, λ_i, of each PCA component. The random GMM parameters are then the multiplicative sum of the Gaussian random variables, $\vec{\xi}$, and the PCA principle components (or eigenvectors) of the GMM model parameters.

2.5. Variations in dose

Given a model of patient-specific variability in respiratory-induced organ motion and dose calculation, we compute statistics of the deposited dose from a single fraction. With the variation in the GMM parameters expressed in terms of the d-dimensional random variable, $\vec{\xi}$, we incorporate the stochastic model of breathing amplitudes into a statistical characterization of the dose distribution, D, resulting from variations in respiratory-induced organ motion. Because the dose distribution is a direct consequence of anatomical configuration that, in turn, is a consequence of breathing amplitude, the random dynamic dose is expressed as $D(\vec{\xi})$.

In our study, we are interested in computing statistics (e.g., mean and variance) on the stochastic dose deposition, $D(\vec{\xi})$. These quantities can help assess the impact of respiratory-induced organ motion variability on SBRT dose distributions.

2.6. Generalized polynomial chaos-stochastic collocation

Determining the behavior of a stochastic system requires that the random inputs of the system be mathematically characterizable stochastic processes (i.e., have a known or estimable underlying distribution). Though Monte Carlo (MC) techniques provide a straightforward means of computing statistics of random fields like $D(\vec{\xi})$, the large number of samples necessary to compute accurate statistics and the significant time to calculate a single dynamic dose deposition renders random sampling Monte Carlo infeasible for clinical use. Several approaches e.g., Latin hypercube sampling (Stein, 1987; Loh, 1996; Helton et al., 2005), the quasi-Monte Carlo method (Morokoff and Caflisch, 1995; Caflisch, 1998; Niederreiter et al., 1998), and the Markov chain Monte Carlo (MCMC) method (Gamerman, 1997; Quian et al., 2003), achieve improved convergence compared to random sampling (or brute-force) Monte Carlo. However, these approaches gain efficiency at the cost of additional restrictions, and none achieve sufficient reduction in sampling size to render computing stochastic dose tractable.

The generalized polynomial chaos-stochastic collocation (gPC-SC) method (Xiu and Hesthaven, 2005; Xiu, 2007) provides a computationally efficient and easily implemented alternative to MC sampling methods, requiring far fewer samples to calculate accurate statistics. Like MC methods, gPC-SC is a sampling method in that it does not require derivation of the stochastic approximating system nor modification of the original deterministic system. In contrast to MC, where the deterministic system (in our case, the forward dose calculation) must be computed at a very large set of randomly chosen sample values of the stochastic input process (the breathing amplitude densities) gPC-SC employs quadrature rules to minimize the number of samples necessary to integrate the stochastic process of interest over the appropriate domain and compute accurate statistics. Under assumptions of smoothness of the system with respect to inputs, which in this case equate to the recognition that the dose distributions vary smoothly as a function of the breathing signal, we gain exponential convergence in the statistical accuracy as a function of the number of dose distribution forward simulations we compute. This process yields a sequence of solutions for a small and far more computationally tractable number of specific realizations of the stochastic field. These solutions are used to obtain highly accurate estimates of the mean, variance, and higher statistical moments of the system.

The generalized polynomial chaos (gPC) method provides a means of representing stochastic processes as a linear combination of orthogonal stochastic polynomials (Xiu and Karniadakis, 2002). In our case, the GMM parameters are Gaussian distributed and can be represented exactly by two Hermite polynomials. Because dose calculation is a non-linear process with respect to the GMM parameters and patient anatomy, the resulting distribution of the dose will be non-Gaussian. Stochastic processes with arbitrary or non-Gaussian distributions are represented using weighted sums of Hermite polynomials as follows: $\vec{\xi}(\omega) = \sum_{i=0}^{N} \alpha_i H_i(\omega)$, where ω is a random variable and α_i is a weight obtained by projecting the stochastic process onto the ith Hermite polynomial.

The Hermite polynomials are given by the recurrence relation:

$$H_{i+1} = 2\omega H_i - 2iH_{i-1}$$
$$H_0 = 1$$
$$H_1 = 2x$$

The stochastic collocation approach consists of selecting a collection of points at which to sample the random field and corresponding weights that account for the underlying stochastic characteristics of the system. Each collocation point, $\vec{\xi}_i$, represents a particular breathing amplitude density for the duration of a treatment selected from the set of likely breathing patterns. We compute the dose deposition for each collocation realization, $D(\vec{\xi}_i)$, using the method described in Section 2.2.

For Gaussian distributed random variables, ψ, of mean zero and unit variance, the collocation points, $\psi_{i,n}$, are the roots of the nth Hermite polynomial and the weights, $c_{i,n}$, are given by $c_{i,n} = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 (H_{n-1}(\psi_i))^2}$.

Though polynomial roots can be approximated using a root-finding method like Newton's method, it is faster to use the Golub–Welsch algorithm (Golub and Welsh, 1969) in the case of Hermite polynomials (Press et al., 1992). We obtain the roots of the nth-order Hermite polynomial by calculating the eigenvalues of the Jacobi matrix, J, composed of the recurrence relation coefficients of the Hermite polynomials, and defined as follows:

$$J_n = \begin{bmatrix} a_0 & \sqrt{b_1} & & & \\ \sqrt{b_1} & a_1 & \sqrt{b_2} & & \\ & \vdots & \vdots & & \\ & & \sqrt{b_{n-2}} & a_{n-2} & \sqrt{b_{n-1}} \\ & & & \sqrt{b_{n-1}} & a_{n-1} \end{bmatrix}. \tag{10}$$

The collocation weights, c_n, are equivalent to the first component of the normalized eigenvectors of the Jacobi matrix J_n (Press et al., 1992). To accommodate Gaussian random variables, $\vec{\xi}$, of arbitrary mean, μ, and variance, σ^2, we map the collocation points as $\vec{\xi}_i = \sigma\psi_i + \mu$. The collocation weights and points can also be extended to multiple stochastic dimensions using tensor products for lower dimensions or the Smolyak construction (Xiu and Hesthaven, 2005; Xiu, 2007) for higher dimensions. We describe the technique in the following section and clearly illustrate the computational savings in Fig. 4.

For each collocation point, $\vec{\xi}_i$, representing a particular breathing amplitude density over the course of a treatment, we calculate the corresponding dose deposition, $D(\vec{\xi}_i)$. The mean and variance of the deposited dose are calculated using the forward dose computations and the collocation weights as follows:

$$\mathbb{E}[D(\vec{\xi})] \approx \sum_{i=0}^{N} c_i D(\xi_i), \tag{11}$$

$$\mathbb{E}[(D(\vec{\xi}) - \mathbb{E}[D(\vec{\xi})])^2] \approx \sum_{i=0}^{N} c_i (D(\xi_i) - \mu(D(\vec{\xi})))^2. \tag{12}$$

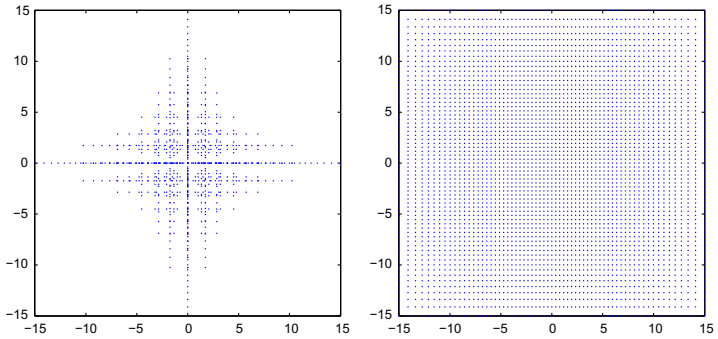


Fig. 4. *The two-dimensional sparse grid* interpolation nodes based on the level 5 Smolyak algorithm and Gauss–Hermite one-dimensional collocation scheme (left) requires only 837 points for the similar integration accuracy of the random process with two independent and uncorrelated Gaussian random variables as the full-tensor product algorithm (right) of the same one-dimensional nodes resulting in 4225 points.

2.7. Smolyak collocation points

For increasing random dimensions, the number of stochastic collocation points necessary to accurately compute integrals grows exponentially. Smolyak's construction (Smolyak, 1963) is a linear combination of one-dimensional tensor product formulas that spans a subspace of the tensor product space and requires far fewer total collocation nodes than the straightforward d-dimensional tensor product of one-dimensional collocation rules.

Given a one-dimensional quadrature rule,

$$Q_i(f) = \sum_{j=1}^{2i+1} c_j f(\omega_j), \tag{13}$$

where ω_j and c_j are the collocation nodes and weights, respectively, the d-dimensional numerical approximation to the integral $\int_\Omega f(\omega)d\omega$ using Smolyak's algorithm is defined recursively as

$$Q_l^d = \sum_{i=0}^{l} (Q_i - Q_{i-1}) \otimes Q_{l-i}^{d-1} \tag{14}$$

where $Q_{-1} = 0$ and $\otimes$ denotes the tensor product of one-dimensional quadrature rules. Alternatively, we can write

$$Q_l^d = \sum_{l-d+1 \leqslant |\vec{i}| \leqslant l} (-1)^{l-|\vec{i}|} \cdot \binom{d-1}{l-|\vec{i}|} \cdot (Q_{i_1} \otimes \cdots \otimes Q_{i_d}), \tag{15}$$

where i is the set of one-dimensional quadrature indices over d dimensions, l is the level of the Smolyak approximation, $|\vec{i}| = \sum_{k=1}^{d} i_k$, and $l \geqslant d$.

The resulting number of sparse-grid collocation points is significantly fewer than for the full-tensor construction providing an accurate cubature formula that does not suffer as significantly from the "curse of dimensionality" (Novak and Ritter, 1997) as the full-tensor construction. This computational savings is clearly illustrated in Fig. 4 which depicts both the full-tensor collocation points (4225 nodes) and the corresponding Smolyak points (837 nodes) for numerical quadrature of a process consisting of two independent and uncorrelated Gaussian random variable inputs. Because we employ three random dimensions in our model, we observe a reasonable savings in computation time with Smolyak versus tensor points. Moreover, we expect significant additional savings when increased model complexity requires the incorporation of a greater number of random dimensions.

3. Results

In this section, we present results for a SBRT liver cancer patient treated with four fractions in the Department of Radiation Oncology at the Huntsman Cancer Institute (HCI). Axial, sagittal, and coronal views of the patient's static dose plan and three of their representative RPM traces recorded on different days are depicted in Fig. 1. The 4DCT images used in this retrospective study were collected at HCI on a 16-slice large bore LightSpeed RT CT scanner (GE Health Care, Waukesha, WI) during the SBRT treatment process using the 4D RCCT (Ford et al., 2003; Vedam et al., 2003) scan protocol described below. Scans at each couch position were continuously acquired in the axial cine mode for a period of time equal to the maximum breathing cycle plus 1 s with a 0.5 s per revolution gantry rotation speed and slice-thickness of 1.25 mm at 120 kVp and GE software slice-thickness optimized mA. A total of roughly 2900 CT slices were acquired at 187 couch positions and the patient's breathing amplitude was continuously recorded during CT acquisition using Varian's RPM system. An additional four RPM respiratory traces were recorded during CT imaging on treatment days and all five traces were subsequently analyzed to determine the variability in patient breathing behavior. The clinical static dose calculations were performed using the BrainSCAN v5.31 (BrainLAB AG, Heimstetten, Germany) radiation treatment-planning (RTP) system's pencil beam algorithm.

Fig. 5 illustrates the two-Gaussian mixture model approximations to the amplitude densities of five recorded RPM breathing traces. Breathing amplitude histograms are depicted to provide a basis for comparing the EM fits to the breathing amplitude data. Panel (f) depicts the GMM fits and clearly illustrates the variations in the breathing amplitude density over the course of several days. Note that both the shape of the amplitude and the absolute values can change significantly. The amplitudes recorded in the first trace (Panel (a)) range from about 3.5 to 4.8 cm, while the amplitudes for the fourth RPM signal (Panel (d)) range from about 4 to 5 cm. Additionally, the qualitative shape of the second, fourth, and fifth (Panels (b), (d), and (e), respectively) GMMs differ greatly slightly from the first and significantly from the third GMMs (Panels (a) and (c), respectively). This variability differs from patient to patient and necessitates patient-specific models of breathing variability and dose delivery uncertainty.

Examination of the eigenvalues corresponding to variation in the parameters depicted in Panel (d) of Fig. 6 suggests that only three PCA components are necessary to accurately capture the variability in breathing. For visual comparison, the reconstruction of the GMM models for the five RPM breathing traces is depicted in

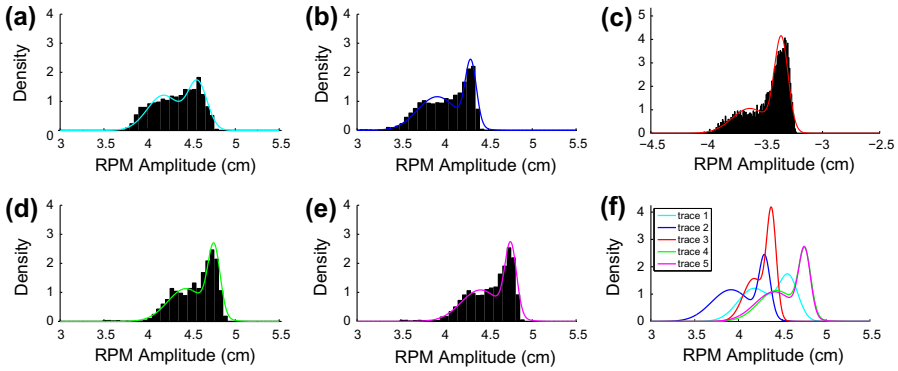


Fig. 5. *The Gaussian mixture model* provides an estimation of amplitude densities of the RPM breathing traces. The GMM fit for each of the five RPM traces overlays the corresponding histogram of breathing amplitudes. The daily variations in amplitude density of the different RPM traces are evident on the bottom right panel.

panels (a), (b), and (c) of Fig. 6. Though the average RMS difference between the GMMs fitted to the breathing and the reconstructed GMMs decreases from 3.8×10^{-4} to 1.7×10^{-16} from three to four components, the eigenvalue of the fourth principal component of the Gaussian Mixture Model parameters is quite small. The additional accuracy gained by including four rather than three components in the reconstruction is on the order of slight variations in the RPM measurement setup and, moreover, does not significantly impact the stochastic dose calculation. As such, it is not sufficient to justify the increased system complexity. The reasonably close correspondence between the original fitted and reconstructed GMMs using only three PCA components enables significant reduction in the complexity of the stochastic system owing to the correspondingly reduced dimensionality of the stochastic space. Thus, we use three components to capture the variation observed in the breathing traces.

Fig. 7 depicts the average and standard deviations of deposited dose over a single treatment for a sagittal view. A comparison of the average dose depositions to the static dose deposition calculation in Fig. 1 shows little difference. However, examination of the standard deviation in dose shows non-trivial high values (greater than 0.2 Gray) occurring near the boundaries of the lesion. From our experiments, we have observed that large standard deviations in dose often correspond to regions of high dose gradient that undergo large respiratory-induced organ deformation. Such areas are significant because they indicate planned dose regions that may differ significantly from actual dose deposition during treatment and are likely candidates for over- or under-dosing.

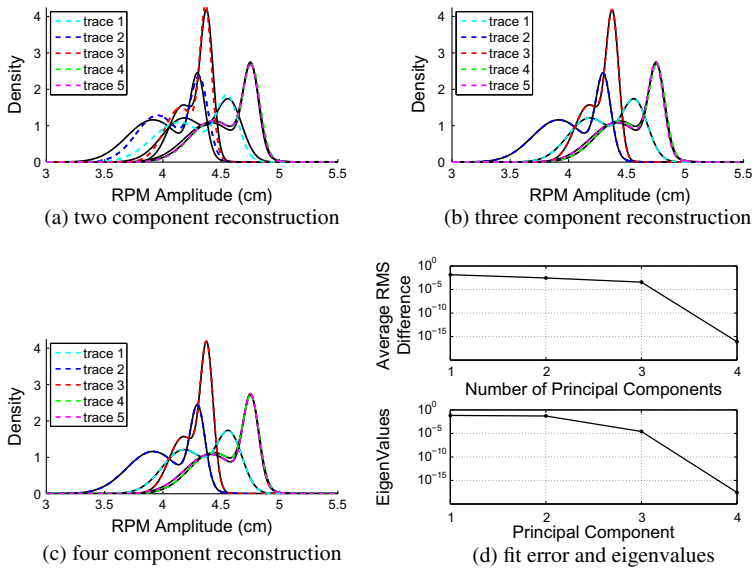


Fig. 6. *The reconstruction of the Gaussian mixture model fits to the breathing traces* are depicted in (a), (b), and (c) for two, three, and four component reconstructions, respectively. The reconstructions are depicted by the dashed color lines and the GMM fits to the patient breathing traces are shown in black. The principal component analysis reconstruction gives very close reproductions of the original mixture model fits with only three independent and uncorrelated eigenvectors (b). Panel (d) depicts the average root mean squared difference between the original fits and the reconstructions for different numbers of principal components in the reconstructions and the eigenvalues corresponding to each component.

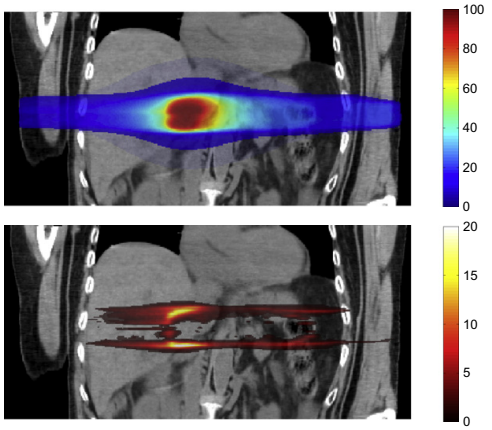


Fig. 7. *The average (top) and standard deviation (bottom) of stochastic dose deposition (in Gray)* is depicted as a percentage of prescribed dose for coronal anatomical views.

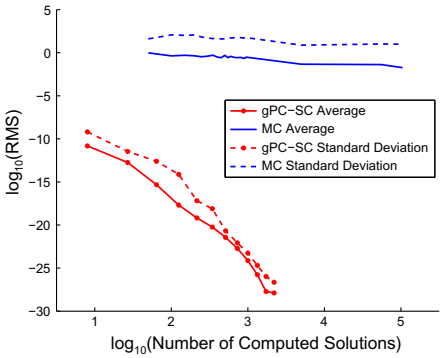


Fig. 8. *The convergence rates for MC and gPC-SC methods* as used to compute average and standard deviation of the deposited dose.

To validate our approach, we present in Fig. 8 the convergence in gPC-SC and traditional MC dose statistics for the patient case depicted in Fig. 7. The convergence data depicted is the RMS difference between the current and final number of forward solutions for the average and standard deviation of dose calculations. It is clear that with only 2744 realizations the gPC-SC method has reached greater convergence than the MC method with 155,000 forward dose solutions. Thus, for this particular model, gPC-SC exhibits significantly faster convergence than MC.

4. Discussion

The goal of this study was to demonstrate the utility and feasibility of a framework for quantifying the variability in respiratory-induced organ motion and incorporate that stochastic model into the calculation of dose deposition for SBRT treatment-planning. In contrast to Monte Carlo methods which are clinically infeasible because they require weeks or even months to compute accurate dose deposition statistics, the efficiency of the proposed approach enables physicians to perform statistical studies of dose response to breathing induced organ motion on a clinically realistic time scale. Statistical dose computations are particularly useful in planning because they allow physicians to identify and avoid dose plans in which high standard deviations in dose coincide with radiation sensitive tissues e.g., the spinal cord and cardiac tissue. We propose that accurate statistical models of predicted dose deposition resulting from organ motion will enable physicians to better assess the impact of SBRT dose plans on normal tissue and tumor lesions and reduce the tumor margins currently incorporated into the clinical SBRT treatment process.

Acknowledgements

This work was funded by a University of Utah Synergy Grant Award (Salter and Joshi), NSF Career Award (Kirby) NSF-CCF0347791, and the National Institute of Biomedical Imaging and Bioengineering Award (Joshi) R01EB007688. The authors would like to acknowledge the computational support and resources provided by the Scientific Computing and Imaging Institute.

References

Abdelnour, A., Nehmeh, S., Pan, T., Humm, J., Vernon, P., Schröder, H., Rosenzweig, K., Mageras, G., Yorke, E., Larson, S., Erdi, Y., 2007. Phase and amplitude binning for 4D-CT imaging. Phys. Med. Biol. 52 (May), 3515–3529.

Babuška, I., Nobile, F., Tempone, R., 2005. A stochastic collocation method for elliptic partial differential equations with random input data. Tech. Rep. ICES 05-47, The University of Texas at Austin.

Beddar, A., Kainz, K., Briere, T., Tsunashima, Y., Pan, T., Prado, K., Mohan, R., Gillin, M., Krishnan, S., 2007. Correlation between internal fiducial tumor motion and external marker motion for liver tumors imaged with 4D-CT. Int. J. Radiat. Oncol. Biol. Phys. 67 (2), 630–638.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision 61 (2), 139–157.

Boldea, V., Sharp, G., Jiang, S., Sarrut, D., 2008. 4D-CT lung motion estimation with deformable registration: quantification of motion nonlinearity and hysteresis. Med. Phys. 35 (3), 1008–1018.

Bortfeld, T., Jokivarsi, K., Goitein, M., Kung, J., Jiang, S., 2002. Effects of intra-fraction motion on IMRT dose delivery: statistical analysis and simulation. Phys. Med. Biol. 47, 2203–2220.

Bortfeld, T., Jiang, S., Rietzel, E., 2004. Effects of motion on the total dose distribution. Semin. Radiat. Oncol. 14, 41–51.

Brandner, E., Wu, A., Chen, H., Heron, D., Kalnicki, S., Komanduri, K., Gerszten, K., Burton, S., Ahmed, I., Shou, Z., 2006. Abdominal organ motion measured using 4D CT. Int. J. Radiat. Oncol. Biol. Phys. 65, 554–560.

Caflisch, R., 1998. Monte Carlo and quasi-Monte Carlo methods. Acta Numerica, vol. 7. Cambridge University Press.

Chi, P., Balter, P., Luo, D., Mohan, R., Pan, T., 2006. Relation of external surface to internal tumor motion studied with CINE CT. Med. Phys. 33 (9), 3116–3123.

Christensen, G.E., Rabbitt, R.D., Miller, M.I., 1996. Deformable templates using large deformation kinematics. IEEE Trans. Image Process. 5 (10), 1435–1447.

Chui, C., Yorke, E., Hong, L., 2003. The effects of intra-fraction organ motion on the delivery of intensity-modulated field with a multileaf collimator. Med. Phys. 30, 1736–1746.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B: Met. 39 (1), 1–38.

Dupuis, P., Grenander, U., Miller, M.I., 1998. Variational problems on flows of diffeomorphisms for image matching. Quart. J. Appl. Math. 56 (3), 587–600.

Ford, E., Mageras, G., Yorke, E., Ling, C., 2003. Respiration-correlated spiral CT: a method of measuring respiratory-induced anatomic motion for radiation treatment planning. Med. Phys. 30, 88–97.

Foskey, M., Davis, B., Goyal, L., Chang, S., Chaney, E., Strehl, N., Tomei, S., Rosenman, J., Joshi, S., 2005. Large deformation three-dimensional image registration in image-guided radiation therapy. Phys. Med. Biol. 50, 5869–5892.

Gamerman, L., 1997. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman & Hall, London, England.

Ganapathysubramanian, B., Zabaras, N., 2007. Sparse grid collocation schemes for stochastic natural convection problems. J. Comput. Phys. 225 (1), 652–685.

Geneser, S., Kirby, R., Wang, B., Salter, B., Joshi, S., 2009. Incorporating patient breathing variations into a stochastic model of dose deposition for stereotactic body radiation therapy. Proceedings of the 21st International Conference on Information Processing in Medical Imaging, vol. 5636. Lecture Notes in Computer Science (LNCS), Williamsburg, Virginia, pp. 688–700.

Golub, G., Welsh, J., 1969. Calculation of Gauss quadrature rules. Math. Comput. 23, A1–A10.

Hanley, J., Debois, M., Raben, A., et al., 1996. Deep inspiration breath-hold technique for lung tumors: the potential value of target immobilization and reduced lung density in dose escalation. Int. J. Radiat. Oncol. Biol. Phys. 36 (1), 188.

Heath, E., Seuntjens, J., Sheikh-Bagheri, D., 2004. Dosimetric evaluation of the clinical implementation of the first commercial IMRT Monte Carlo treatment planning system at 6 MV. Med. Phys. 31, 2771–2779.

Helton, J., Davis, F., Johnson, J., 2005. A comparison of uncertainty and sensitivity analysis results obtained with random and Latin hypercube sampling. Reliab. Eng. Syst. Safe. 89, 305–330.

Herk, M.v., 2004. Errors and margins in radiotherapy. Semin. Radiat. Oncol. 14 (1), 52–64.

Hinkle, J., Fletcher, P.T., Wang, B., Salter, B., Joshi, S., 2009. 4D MAP image reconstruction incorporating organ motion. In: IPMI 2009: Proceedings of Information Processing in Medical Imaging, pp. 676–687.

Ionascu, D., Jiang, S., Nishloka, S., Shirato, H., Berbeco, R., 2007. Internal–external correlation investigations of respiratory induced motion of lung tumors. Med. Phys. 34 (10), 3893–3903.

Jiang, S., Pope, C., Al Jarrah, K., Kung, J., Bortfeld, T., Chen, G., 2003. An experimental investigation on intra-fractional organ motion effects in lung imrt treatments. Phys. Med. Biol. 48, 1773–1784.

Jiang, S., Wolfgang, J., Mageras, G., 2008. Quality assurance challenges for motion-adaptive radiation therapy: gating, breath holding, and four-dimensional computed tomography. Int. J. Radiat. Oncol. Biol. Phys. 71 (1), S103–S107.

Joshi, S.C., Miller, M.I., 2000. Landmark matching via large deformation diffeomorphisms. IEEE Trans. Image Process. 9 (8), 1357–1370.

Keall, P., Todor, A., Vedam, S., Bartee, C., Sibers, J., Kini, V., Mohan, R., 2004. On the use of EPID-based implanted marker tracking for 4D radiotherapy. Med. Phys. 31, 3492–3499.

Keall, P., Joshi, S., Vedam, S., Siebers, J., Kini, V., Mohan, R., 2005a. Four-dimensional radiotherapy planning for DMLC-based respiratory motion tracking. Med. Phys. 32, 942–951.

Keall, P., Kini, V., Vedam, S., Mohan, R., 2005b. Potential radiotherapy improvements with respiratory gating. Australas. Phys. Eng. Sci. Med. 25, 1–6.

Keall, P., Mageras, G., Balter, J., Emery, R., Forster, K., Jiang, S., Kapatoes, J., Low, D., Murphy, M., Murray, B., Ramsey, C., Van Herk, M., Vedam, S., Wong, J., Yorke, R., 2006. The management of respiratory motion in radiation oncology report of AAPM Task Group (76^a). Med. Phys. 33, 3874–3900.

Kubo, H., Len, P., Minohara, S., Mostafavi, H., 2000. Breathing-synchronized radiotherapy program at the University of California Davis Cancer Center. Med. Phys. 27 (2), 346–353.

Kuo, H.-H., 1975. Gaussian Measures in Banach Spaces. Lecture Notes in Mathematics, vol. 463. Springer-Verlag, Berlin, Heidelberg, New York.

Loh, W., 1996. On Latin hypercube sampling. Ann. Stat. 24, 2058–2080.

Lujan, A., Larsen, E., Balter, J., Ten Haken, R., 1999. A method for incorporating organ motion due to breathing into 3D dose calculations. Med. Phys. 26 (5), 715–720.

Lujan, A., Larsen, E., Balter, J., Ten Haken, R., 2003. A method for incorporating organ motion due to breathing into 3D dose calculations: sensitivity to variations in motion. Med. Phys. 10, 2643–2649.

Mathelin, L., Hussaini, M., 2003. A stochastic collocation algorithm for uncertainty analysis. Tech. Rep. NASA/CR-2003-212153, NASA Langley Research Center.

McGarry, R., Papiez, L., Williams, M., Whitofrd, T., Timmerman, R., 2005. Stereotactic body radiation therapy of early-stage non-small-cell lung carcinoma: phase I study. Int. J. Radiat. Oncol. Biol. Phys. 64, 1010–1015.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, Inc., New York, NY.

Morokoff, W., Caflisch, R., 1995. Quasi-Monte Carlo integration. J. Cell Physiol. 122 (2), 218–230.

Naqvi, S.A., D'Souza, W.D., Yu, C., 2005. Real-time intra-fraction-motion tracking using the treatment couch: a feasibility study. Phys. Med. Biol. 50, 4021–4033.

Nehmeh, S., Erdi, Y., Pan, T., Yorke, E., Mageras, G., Rosenzweig, K., Schoder, H., Mostafavi, H., Squire, O., Pevsner, A., et al., 2004. Quantitation of respiratory motion during 4D-PET/CT acquisition. Med. Phys. 31, 1333–1338.

Neicu, T., Berbeco, R., Wolfgang, J., Jiang, S., 2006. Synchonized moving aperture radiation therapy (SMART): improvement of breathing pattern reproducibility using respiratory coaching. Phys. Med. Biol. 51, 617–636.

Niederreiter, H., Hellakalek, P., Larcher, G., Zinterhof, P., 1998. Monte Carlo and Quasi-Monte Carlo Methods. Springer-Verlag, New York, NY.

Nobile, F., Tempone, R., Webster, C., 2008. A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. 46 (5), 2309–2345.

Novak, E., Ritter, K., 1997. The curse of dimension and a universal method for numerical integration. In: Nurnberger, G., Schmidt, J., Walz, G. (Eds.), Multivariate Approximation and Splines, pp. 177–187.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2 (6), 559–572.

Pevsner, A., Davis, B., Joshi, S., et al., 2006. Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated CT images. Med. Phys. 33 (2), 369–376.

Press, W., Teukolsky, S., Vetterling, W., Flannery, B., 1992. Numerical Recipes in C: The Art of Scientific Computing, second ed. Cambridge University Press, New York, NY (Chapter: Gaussian Quadratures and Orthogonal Polynomials, pp. 147–161).

Purdy, J., 2001. Intensity-modulated radiotherapy: current status and issues of interest. Int. J. Radiat. Oncol. Biol. Phys. 51 (4), 880–914.

Quian, S., Stow, C., Borsuk, M., 2003. On Monte Carlo methods for Bayesian inference. Ecol. Modell. 159, 269–277.

Rassiah-Szegedi, P., Salter, B., Fuller, C., Blough, M., Papanikolaou, N., Fuss, M., 2006. Monte Carlo characterization of target doses in stereotactic body radiation therapy (SBRT). Acta Oncol. 45, 989–994.

Sawant, A., Venkat, R., Srivastava, V., Carlson, D., 2008. Management of three-dimensional intrafraction motion through real-time DMLC tracking. Med. Phys. 35 (5), 2050–2061.

Siantar, C.H., Walling, R., Daly, T., Faddegon, B., Albright, N., Bergstrom, P., et al., 2001. Description and dosimetric verification of the PEREGRINE Monte Carlo dose calculation system for photon beams incident on a water phantom. Med. Phys. 28, 122–1337.

Smolyak, S., 1963. Quadrature and interpolation formulas for tensor products of certain classes of functions. Soviet Math. Dokl. 4, 240–243.

Stein, M., 1987. Large sample properties of simulations using Latin hypercube sampling. Technometrics 29, 143–151.

Timmerman, R., Forster, K., Chinsoo, C., 2005. Extracranial stereotactic radiation delivery. Semin. Radiat. Oncol. 15, 202–207.

Vedam, S., Keall, P., Kini, V., Mostafavi, H., Shukla, H., Mohan, R., 2003. Acquiring a four-dimensional computed tomography dataset using an external respiratory signal. Phys. Med. Biol. 48 (1), 45–62.

Vedam, S., Docef, A., Fix, M., Purphy, M., Keall, P., 2005. Dosimetric impact of geometric errors due to respiratory motion prediction on dynamic multileaf collimator-based four-dimensional radiation delivery. Med. Phys. 32, 1607–1620.

Wiener, N., 1938. The homogeneous chaos. Am. J. Math. 60 (4), 897–936.

Wijesooriya, K., Weiss, E., Dill, V., Dong, L., Mohan, R., Joshi, S., 2008. Quantifying the accuracy of automated structure segmentation in 4D CT images using a deformable image registration algorithm. Med. Phys. 35 (4), 1251–1260.

Wu, Q., Thongphiew, D., Wang, Z., Chankong, V., Yin, F., 2008. The impact of respiratory motion and treatment technique on stereotactic body radiation therapy for liver cancer. Med. Phys. 35 (4), 1440–1451.

Xing, L., Thorndyke, B., Schreibmann, E., Yang, Y., Li, T., Kim, G., Luxton, G., Koong, A., 2006. Overview of image-guided radiation therapy. Med. Dosim. 31 (2), 91–112.

Xiu, D., 2007. Efficient collocational approach for parametric uncertainty analysis. Commun. Comput. Phys. 2 (2), 293–309.

Xiu, D., Hesthaven, J., 2005. High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. 27 (3), 1118–1139.

Xiu, D., Karniadakis, G., 2002. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. 24, 619–644.